



## DEPARTMENT OF DEFENSE

POLYGRAPH INSTITUTE  
7540 PICKENS AVENUE  
FORT JACKSON, SOUTH CAROLINA 29207

October 7, 2002

MEMORANDUM FOR DEFENSE TECHNICAL INFORMATION CENTER, 8725 JOHN  
KINGMAN ROAD, SUITE 0944, FORT BELVOIR, VIRGINIA  
22060-6218

SUBJECT: Report Submission

The Department of Defense Polygraph Institute (DoDPI) submits the following report, Scaled P300 Scalp Profiles In Detection Of Deception (DoDPI02-R-0005), for inclusion to your collection of scientific and technical information for the Department of Defense (DoD) community.

The DoDPI point of contact for this action is Rose Swinford, DSN 734-9163,  
commercial (803) 751-9163

*William F. Norris*  
WILLIAM F. NORRIS  
Director

2 Attachments

1. SF 298 – Report Documentation Page
2. Report

# REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE September 2002	3. REPORT TYPE AND DATES COVERED October 1997 - August 2002
4. TITLE AND SUBTITLE Scaled P300 Scalp Profiles in Detection of Deception			5. FUNDING NUMBERS DoDPI98-P-0001
6. AUTHOR(S) J. Peter Rosenfeld, Ph.D.			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Northwestern University Department of Psychology Evanston, IL 60208			8. PERFORMING ORGANIZATION REPORT NUMBER DoDPI02-R-0005
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) DoD Polygraph Institute 7540 Pickens Avenue Fort Jackson, SC 29207			10. SPONSORING / MONITORING AGENCY REPORT NUMBER DoDPI02-R-0005
11. SUPPLEMENTARY NOTES Public Release, Distribution Unlimited			
12a. DISTRIBUTION / AVAILABILITY STATEMENT			12b. DISTRIBUTION CODE
13. ABSTRACT (Maximum 200 words) Three studies were performed. The first two dealt with countermeasures to brain wave-based detection of deception in concealed information test protocols. There are two kinds of such protocols extant. One, the "6-probe" protocol utilizes multiple different crime details whose brain responses are averaged together. This protocol was easily defeated in the first study, as the detection rates dropped from 82% detection in the simple guilty group to 18% in the guilty group using a countermeasure. Although the average reaction time distinguished these two groups, there was enough overlap in their reaction time distributions such that in any individual case, one could not use reaction time to infer deception. The second protocol, the "1-probe" protocol uses one crime detail as a probe in each of as many runs as one wishes. One group was run in three successive weeks as 1) a guilty group, 2) a countermeasure group, and 3) finally without the explicit use of the countermeasure. In the first week, 92% of the subjects were detected. The countermeasure dropped this rate to 50%. In the final third week, without explicit use of the countermeasure, only 58% were detected. There was no overlap in the reaction time distributions of the first two weeks, suggesting that the explicit countermeasure use could be detected with reaction time. In the third week, the reaction time distributions looked like those of the first week, so that test beaters would not be detected with reaction time. Other matters examined were 1) a comparison of individual brain wave analysis methods; 2) a comparison of naive versus sophisticated subjects, and 3) a comparison in terms of workload between the 1-probe and the 6-probe protocols			
14. SUBJECT TERMS Psychophysiological detection of deception, P300, screening tests, event-related potentials, brain maps			15. NUMBER OF PAGES 98
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT U	18. SECURITY CLASSIFICATION OF THIS PAGE U	19. SECURITY CLASSIFICATION OF ABSTRACT U	20. LIMITATION OF ABSTRACT None

Running head: SCALED P300 SCALP PROFILES IN DETECTION OF DECEPTION

Report No. DoDPI02-R-0005

Scaled P300 Scalp Profiles In Detection Of Deception

J. Peter Rosenfeld, Ph.D.

Northwestern University

Department of Psychology

Evanston, IL 60208

September 2002

Department of Defense Polygraph Institute  
Fort Jackson, South Carolina 29207

### Acknowledgements

Funds for this research were provided by the Department of Defense Polygraph Institute under project number DODPI98-P-0001. The views expressed in this article are those of the author and do not reflect the official policy or position of the department of Defense or the U.S. Government.

## Executive Summary

Since 1987, researchers have been experimenting with brain waves as a means of detecting deception. The most frequently used brain wave is called P300. This is one of a class of Event-Related or Evoked potentials, "ERPs". To record ERPs, one puts brain wave ("EEG") electrodes on the scalp, but instead of looking at spontaneous brain activity, one looks at the series of peaks and troughs following particular stimuli. If one presents to a subject a series of dates, one at a time, every 3 or 4 seconds, occasionally presenting the subject's birth date, the P300 wave will appear following these rare and personally meaningful birth date presentations. To better see the P300, averages are made of all the single ERP birthday responses, and these are superimposed upon an average of the frequent irrelevant responses. The former average will ordinarily contain a large P300 wave, a positive-going deflection which begins 300 to 600 milliseconds(ms) after the stimulus and returns to baseline about 500 ms later. The irrelevant average will be a relatively flat line.

It may be evident that this protocol can be used in a "Guilty Knowledge Test" (GKT, also called a concealed information test or CIT). One typically uses an actual detail from a crime scene as the rare event (called a "probe"), and then, several plausible but fictitious details as the frequent (called "irrelevant") stimuli. The idea is that only the perpetrator and the authorities recognize the actual criminal detail, so that only the perpetrator will generate a P300 in response to it. The same reservations one may pose about the polygraphic CIT apply to the ERP version. For example, will a criminal remember the critical detail which the test giver believes should be remembered? Then there is the question of countermeasures. Until the present project, there has been no research about the resistance of these ERP-based CITs to countermeasures.

It may be of concern that GKTs (CITs) are not of interest to the government which mostly utilizes Comparison Question Tests (CQTs). It should be immediately pointed out that when P300-based tests are used, the brain wave responses are readily adaptable to comparison question analogs, including screening tests, and the present investigator published such applications in 1991 and 1992. Thus what one learns with a P300-based GKT will likely apply to a P300-based CQT.

There have been two versions of this ERP-based CIT. One combines a set of 6 ERP responses to 6 different probes (each repeated multiple times) within one probe average. Likewise, there are 4 irrelevant stimuli per probe, i.e., 24 irrelevants repeated multiple times. There is one more stimulus used in this protocol, called a target. Subjects are told to press a yes button to the targets and a no button to all other stimuli. There are 6 targets, one per probe, all in the same category of that probe. So if a stolen item is a ring, that is one probe, and the target could be bracelet, and the irrelevants would be 4 other pieces of (not stolen) jewelry; watch, broach, necklace, tiara. The target forces the subject to pay attention to stimuli, but also elicits a benchmark P300 as a rare, meaningful (the only item to which one presses "yes") stimulus. The logic of this test is that if a subject is guilty, the probe and the target will elicit P300. If the subject is innocent, only the target elicits the P300. Thus in a guilty subject, the responses to probe and *target* should look alike, but in an innocent subject, the probe and the *irrelevant* should look alike. The analytic method used to make this determination is to calculate and compare cross correlations of probe and target waveforms versus cross-correlations of probe and irrelevant waveforms. The cross-correlation simply aligns the waveforms in question and determines quantitatively the goodness of alignment. This is called the BC-AD test in the full report. The studies done using this protocol have typically been done on

advanced cooperative paid volunteers, and friends or associates of the investigators. Only one scalp site, called Pz, is recorded. A mock crime scenario is typically used.

The other protocol for this ERP-based CIT uses the same kind of probe and irrelevant stimuli, but the target is optional since 1) the main comparison is between probe and irrelevant P300 size (the difference should be large in guilty subjects, otherwise small), and 2) there are other ways of forcing attention besides using targets. This protocol simply locates and finds the P300 peak (maximum positive) amplitude within a time window (usually 450-950 ms) *wherever it is* for both probe and irrelevant responses and statistically compares them. In the report, this test is called the BAD test. Another crucial difference is that in this protocol, for each run there is only one probe and 6 or more irrelevant stimuli, all stimuli repeated multiple times. This protocol works with the most naive subjects as well as with advanced subjects. Here too, one usually uses data from the single Pz scalp site.

We developed similar countermeasures for each protocol. The idea was to make the irrelevant relevant by assigning covert responses, mental and physical, to presentations of irrelevant stimuli. If done right, this procedure should produce PSOs in irrelevant responses which will now resemble probe and target responses, defeating the tests. In the first "6-probe protocol," subjects were taught to recognize the categories of stimuli—such as the stolen jewelry category—and generate one kind of covert response for each category. For the "1-probe protocol," subjects were taught to generate specific covert responses for each irrelevant. We also ran and compared results with advanced versus naive subjects (since they are more likely the type of subject one will run into in the field).

#### Main Results, 6-probe paradigm

- 1) In a naive *guilty* group with no countermeasure (CM), the BAD method detected 82% of the subjects but the BC-AD method detected only 54%. In a later follow-up study with sophisticated subjects, both analytic methods performed at about 85% accuracy. In an *innocent* group, the false positive rate was 0 (BC-AD) to 9% (BAD). In the CM group, the BAD method detected only 18% of the subjects, and the BC-AD method continued to detect a chance 54%. **One concludes that this protocol is vulnerable to countermeasures, and that BAD is a better diagnostic approach than BC-AD.**
- 2) One might have expected that reaction time (RT) would have helped to identify CM-users, as the RTs to irrelevant stimuli would be expected to increase as subjects thought on each irrelevant trial about which covert response to make. On average this was so, but there was considerable overlap of RT distributions between simple guilty and CM groups for all stimulus types, so that **in any given individual, the test can be beaten without a hint from reaction time data of CM use.**

#### Main Results, 1-probe paradigm.

- 1) In this study, we ran the same group of advanced subjects through 3 weeks of tests, the first week with no CM, the second, with the CM, and the third, without the CM. In the first week, BAD detected 92% of the 13 subjects, BC-AD detected 69%. In the second week with the explicit CM in use, BAD detected only 50%, BC-AD detected 25%. The explicit CM worked. Surprisingly and interestingly, in the third week, without the explicit CM in use, BAD detected only 58% of the subjects who were not even trying to beat the test. Moreover, in the third week, the subjects who beat the test were indistinguishable in reaction time or ERPs from genuinely innocent subjects.

- 2) During the second week, when the subjects were explicitly using the CM, there was *no overlap* of the RT distribution to irrelevant stimuli with the irrelevant RT distribution of the first week in the naive guilty subjects. **This suggests that RT can reveal explicit CM use in the 1-probe protocol. But in that third week, in those 42% of subjects still undetected, RTs are no help and these subjects totally defeat the test.**

### Conclusions and introduction to final scalp distribution studies.

It is evident that the ERP protocols investigated above are seriously vulnerable to CMs, and in view of other weaknesses of the CIT/GKT approach, a novel protocol was developed in which the brain map of P300 amplitudes across the scalp was utilized as a novel response channel for detection of deception. This was the research originally proposed and funded. It involved recording the voltage from 30 electrodes, and comparing the *scaled* scalp distributions of P300 amplitudes across these 30 electrodes to relevant and control questions in a CQT screening analog. (Scaling renders the brain map orthogonal to simple amplitude. That is, the map shape yields information not correlated to the size of the brain waves.) To compare scalp distributions *within subjects* to differing stimuli, it was necessary to develop wholly novel analytic methods.

In this study the relevants were like probes, and the controls were innocent acts, somewhat comparable to irrelevant. There was no mock crime in these studies; rather, there were 8 anti-social/illegal acts that the subjects were asked about on a display screen, one at a time every 4 seconds. They were to deny all wrongdoing, but to respond "yes" to a target stimulus. We accused them of 4 acts from the list of acts they looked at prior to the run. Some of these acts were falsely accused, some truly. Ground truth was obtained after the run by having them check items on a list of acts in perceived privacy while their lists were secretly observed via closed circuit TV. (Of course they were fully de-briefed later). To reduce the dimensionality of the data set, a principal component analysis (PCA—this is like a factor analysis) was done across the dimension of scalp surface so as to identify a set of virtual sites, each including a cluster of actual sites containing redundant information. Each virtual site, however, was orthogonal to the others, that is, each virtual site yielded information not correlated with information from other virtual sites. We ultimately utilized 4 virtual sites accounting for more than 73% of the variability in the entire data set.

### Results:

- 1) We ultimately analyzed data from 15 guilty and 7 innocent subjects. Our most sensitive analysis routine (called CAT in the report) for comparing relevant and control distributions yielded 73% correct detection of guilty subjects and 0% false positives in innocent subjects, both held to the same statistical criteria.
- 2) Using the previous BAD analysis (relevant versus control) on the single virtual site 1, we detected 80% of the guilty subjects, with 0% false positives. The 8 subjects guilty of 2 items were detected as readily as the 7 subjects guilty of 1 item. This is the first time this has been seen.
- 3) Our judgements (about interaction presence or absence) based on visual inspection of graphed scalp distributions within subjects was confirmed by the results of the statistical (CAT) results, lending some validity to these novel methods.

Conclusions:

The scalp distribution method has promise. It should be improved by follow up studies because there is no obvious countermeasure as there is with the simple P300 amplitude tests. This is because there is general knowledge of the determinants of P300 amplitude at one site. There is no knowledge of the determinants of the shapes of scalp distributions.

## Abstract

ROSENFELD, J. P. Scaled P300 Scalp Profiles in Detection of Deception. September 2002, Report No. DoDPI02-R-0011. Department of Defense Polygraph Institute, Fort Jackson, SC 29207-5000.--

Three studies were performed. The first two dealt with countermeasures to brain wave-based detection of deception in concealed information test protocols. There are two kinds of such protocols extant. One, the "6-probe" protocol utilizes multiple different crime details whose brain responses are averaged together. This protocol was easily defeated in the first study, as the detection rates dropped from 82% detection in the simple guilty group to 18% in the guilty group using a countermeasure. Although the average reaction time distinguished these two groups, there was enough overlap in their reaction time distributions such that in any individual case, one could not use reaction time to infer deception. The second protocol, the "1-probe" protocol uses one crime detail as a probe in each of as many runs as one wishes. One group was run in three successive weeks as 1) a guilty group, 2) a countermeasure group, and 3) finally without the explicit use of the countermeasure. In the first week, 92% of the subjects were detected. The countermeasure dropped this rate to 50%. In the final third week, without explicit use of the countermeasure, only 58% were detected. There was no overlap in the reaction time distributions of the first two weeks, suggesting that the explicit countermeasure use could be detected with reaction time. In the third week, the reaction time distributions looked like those of the first week, so that test beaters would not be detected with reaction time. Other matters examined were 1) a comparison of individual brain wave analysis methods; 2) a comparison of naive versus sophisticated subjects, and 3) a comparison in terms of workload between the 1-probe and the 6-probe protocols.

The third study looked at brain maps in detection of deception. In it, twenty-three subjects had P300 event-related brain potentials (ERPs) from 30 scalp sites recorded in response to relevant, control, irrelevant, and target questions as in screening. They also gave behavioral responses. A Principal Component analysis (like a factor analysis) was performed on the sites so as to reveal clusters (principal components) of correlated sites. Four clusters were found to account for 74% of the total variance in all averaged ERPs. Sites within clusters were then averaged to yield virtual sites on which analyses were performed. We compared scalp distributions to relevant versus control stimuli and found that 73% of the subjects guilty of either one (n=7) or two (n=8) items were correctly identified using criteria which did not produce any false positives in seven innocent subjects. We also saw that simple amplitude at the first principal component was greater for guilty responses than for control responses. Principal component clusters were more sensitive in this regard than individual sites. It is recommended that the distribution approach be further investigated since it is probably less vulnerable to a countermeasure than simple amplitude.

**Key Words:** Psychophysiological detection of deception, P300, screening tests, event-related potentials, brain maps.

## Table of Contents

Title Page .....	i
Acknowledgements .....	ii
Executive Summary .....	iii
Abstract .....	vii
Table of Contents .....	viii
Introduction .....	1
Countermeasure Studies: Introduction .....	2
Countermeasure Study 1, Introduction, Methods .....	4
Countermeasure Study 1, Expectations, Results .....	7
Countermeasure Study 1, Discussion, Introduction to Study 2 .....	17
Countermeasure Study 2: Methods, Expectations .....	20
Countermeasure Study 2: Results .....	21
Countermeasure Studies: General Discussion .....	31
Scalp Distribution Studies: Introduction .....	33
Scalp Distribution Studies: Methods .....	35
Scalp Distribution Studies: Results .....	45
Scalp Distribution Studies: Expectations .....	47
Scalp Distribution Studies: Validity of Methods .....	57
Scalp Distribution Studies: Discussion .....	58
References .....	61
Appendix 1 (A publication) .....	64
Appendix 2 (Instructional scripts, stimulus materials, scenarios) .....	86

## INTRODUCTION

### **Background**

We and others pioneered the use of P300 amplitude at one site as an index of recognition in a guilty knowledge (GKT) paradigm (Rosenfeld et al., 1987, 1988; Farwell & Donchin, 1991; Allen, Iacono, & Danielson, 1992). P300 is one of several components of the event-related potential (ERP) seen in the electroencephalogram (EEG) in response to rare, meaningful stimuli (Fabiani, Gratton, Karis, & Donchin, 1987). We also extended this method to a screening analog (Rosenfeld, Angel, Johnson, & Quian, 1991). We have continued examining this paradigm more recently (Ellwanger, Rosenfeld, Sweet, & Bhatt, 1996; Ellwanger, Rosenfeld, Hankin, & Sweet, 1999), but have also switched directions to study another brain-wave measure (the *profile*, described below) for these reasons: 1) P300 amplitude may help in deception detection but is not a direct/specific index of deception; it is an index of recognition, although one may sometimes infer deception when a P300 is emitted in response to a stimulus which the subject denies recognizing. 2) In laboratory situations, 85-90% detection rates are typically reported (Rosenfeld et al., 1991; Farwell & Donchin, 1991; Allen et al., 1992). In the only published field test of this method in a peer-reviewed journal (Miyake, Mizutani, & Yamahura, 1993), the detection rate for guilty subjects was a disappointing 48%. 3) We have very recently developed countermeasures for this P300-based GKT which dramatically reduced its accuracy in a laboratory situation. These studies, although clearly related to P300-based deception detection research, were not originally specifically proposed, but were done during the grant period in the PI's lab and were inevitably benefited by the award, since some of the equipment bought with the award was shared by these studies. It is thus appropriate to report on them here. (Preliminary countermeasure studies were included in the appendix of the original report, and peer reviewers of that report wanted them not in the appendix, but in the main text.) These studies are based on the concealed information test (CIT) paradigm (sometimes called the "guilty knowledge test" or GKT) which the DODPI has not favored recently in deference to the comparison question test (CQT). Nevertheless, they are presented here because in the post 9/11 climate in the U.S., such methods have been heavily promoted as having application to counter-terrorism (<http://www.brainwavescience.com/> which is the web site of Dr. L. Farwell). Although one may question claims of this particular promoter (as the present writer often has; e.g. USGAO, 2001), the logic of using this CIT in counter-terrorism has good face value and is hard to dispute. Moreover there is *laboratory* evidence in the above cited web site and in Farwell & Smith (2001) that the method may be effective--*in the absence of countermeasures (CMs)*, which is of course why the method should be challenged with countermeasures. The reasoning underlying the use of the P300 test in counter-terrorism is as follows: A genuine terrorist will recognize signals, acronyms, initials, colleague names, etc. frequently used in the conversation of the terrorist organization; an innocent suspect will lack this knowledge. Thus, such items, when used in the P300 oddball paradigm, will be recognized by the terrorist (not by the innocent person) and evoke a P300 which will give him away. The countermeasure data next presented will show that the approach is logical, but subject to defeat with a simple countermeasure.

Finally, with respect to the concern that GKTs (CITs) are not of interest to the government which mostly utilizes Comparison Question Tests (CQTs). It should be pointed out that when P300-based tests are used, the brain wave responses are readily adaptable to comparison question analogs, including screening tests, and the present investigator published

such applications in 1991 and 1992. Thus what one learns with a P300-based GKT will likely apply to a P300-based CQT.

# 1. COUNTERMEASURE STUDIES: Introduction.

In the studies of P3 amplitude as a recognition index for concealed information, there are typically three kinds of stimuli presented to subjects, 1) Probes (P), which concern concealed information known only to guilty persons and authorities, 2) Irrelevants (I), which are items irrelevant to the interests of authorities and unrelated to criminal acts, and 3) Targets (T), which are items to which subjects are asked to press a “yes” button. In this report and in previous studies, Ps and Ts have a 1/6 probability, Is have a 4/6 probability. The items are randomly presented one at a time on a video display screen every three seconds (as recommended by Farwell & Smith, 2001). The dishonest subject will press a “no” button to each P occurrence, falsely signaling non-recognition. He will press the “no” button honestly to the irrelevant stimuli, and the “yes” button honestly and as instructed to the T stimuli. The T stimuli serve two purposes: First, they force the subject to attend to the display, as failure to respond appropriately to T stimuli will suggest non-cooperation. Second, the T is a rare and task-relevant stimulus which evokes a benchmark P3 with which other ERPs can be compared in some analyses (described below).

The basic assumption of the P300 CIT is that the P is recognized (though verbally denied) by the dishonest subject, and is thus a rare but meaningful stimulus capable of evoking P300. For the innocent subject, the P is simply another I and should evoke no P300 or a small one. There have been two analytic approaches taken in order to diagnose guilt or innocence. Ours (e.g., Soskins et al., 2001) has been to compare P and I responses; in guilty subjects, one expects  $P > I$ . We use what is described in the next paragraph as the Bootstrapped Amplitude Difference method. The other approach, introduced by Farwell & Donchin, (1991), is based on the expectation that in guilty persons, the P and T stimuli should evoke similar P3 responses, whereas in the innocent subject, P responses will look more like I responses. Thus in this approach, called here Bootstrapped Correlation Analysis of Disparity, the cross correlation of P and T is compared with that of P and I. In guilty subjects, the PT correlation  $>$  PI correlation. The opposite is expected in innocents.

## **Bootstrapped Amplitude Difference (BAD)**

The first technique used to compare ERP responses to stimuli within individuals is called the bootstrapped amplitude difference method (BAD). It observes and compares only amplitude of ERP components and does not consider scalp distribution. To determine whether or not the P300 evoked by one stimulus is greater than that evoked by another within an individual, the bootstrap method is usually used on sites separately (See tutorial by Wasserman & Bockenholt, 1989, for in depth theory and discussion). This will be illustrated with an example of a probe response being compared with an irrelevant response. More specifically, the question being answered by the bootstrap method is: “Is the probability more than 95 in 100 (or 90 in 100) that the true difference between the average probe P300 and the average irrelevant P300 is greater than zero?” For each subject, however, one has available only one average probe P300 and one average irrelevant P300. Answering the statistical question requires distributions of average P300 waves, and these actual distributions are not available. One thus bootstraps the distributions, in the bootstrap variation used here, as follows: A computer program goes through

the probe set and draws at random, with replacement, a set of  $n_1$  waveforms. It averages these and calculates P300 amplitude from this single average using the maximum segment selection methods as described above. Then a set of  $n_2$  waveforms is drawn randomly with replacement from the irrelevant set, from which average P300 amplitude is calculated. The number  $n_1$  is the actual number of accepted probe sweeps for that subject, and  $n_2$  is the actual number of accepted irrelevant sweeps for that subject. The calculated irrelevant mean P300 is subtracted from the comparable probe value, and one thus obtains a difference value to place in a distribution which will contain 100 values after 100 iterations of the process just described. Iterations will yield differing (variable) means and mean differences due to the sampling-with-replacement process.

In order to state with 95% confidence that probe and irrelevant evoked ERPs are indeed different, one requires that the value of zero difference not be within 1.65 SDs from the mean of the distribution of differences, (1.29 for 90% confidence). It is noted that sampling different numbers of probes and irrelevants could result in differing errors of measurement, however, studies have shown a false positive rate of zero utilizing this method (Ellwanger et al., 1996) and others have taken a similar approach (Farwell & Donchin, 1991) with success. This method has the advantage of utilizing all the data, as would an independent groups t-test with unequal numbers of subjects. It is further noted that a one-tailed 1.65 criterion yields a  $p < .05$  confidence level because the hypothesis that the probe evoked P300 is greater than the irrelevant evoked P300 is rejected either if the two are not found significantly different or if the irrelevant P300 is found larger.

#### Bootstrapped Correlation Analysis of Disparity (BC-AD)

The other analysis method to be used to compare ERPs within individuals is called the bootstrapped correlation-analysis of disparity (BC-AD). BC-AD determines if 90% or more of the 100 iterated, double centered cross correlation coefficients between ERP responses to probe and target stimuli are greater than the corresponding cross correlations of responses to the probe and irrelevant stimuli. If so, the subject is found to be guilty (this is the Farwell & Donchin, 1991 criterion and method). For example, within each subject, the program starts with all sweeps to probe (P,  $n=50$ ), target (T,  $n=50$ ), and irrelevant (I,  $n=200$ ) stimuli, and, at each time point, determines the overall, usual within subject ERP average. This series of points comprising the average ERP is called **A**, (a vector). Then the computer randomly draws, with replacement, 50 P sweeps from the P sample of 50. These are averaged to yield a bootstrapped P average. Similarly, a bootstrapped T-average, and I-average are obtained, except the latter is based on a 200-size draw. **A** is now subtracted from P, T, and I. This is double centering and is performed because it enhances the differences among ERP responses. The first of 100 Pearson cross correlation coefficient pairs are now computed for the cross correlation of P and T and for P and I ( $R_{pt}$  and  $R_{pi}$ , respectively). The difference between these two R-values,  $D_1$ , is computed. The process is iterated 100 times yielding  $D_2, D_3 \dots D_{100}$ . Using the Farwell and Donchin (1991) method, the number of D-values in which  $R_{pt} > R_{pi}$  is then counted. If this number is greater than or equal to 90, a guilty decision is made. If this number is less than 90, then a not guilty decision is made. In the presented studies, a mathematically similar criterion is used in which the distribution of D values is considered. If zero is more than 1.29 SDs (90% confidence) below the mean, then a guilty decision is made.

Although we have not yet spoken of Reaction Time (RT) in these studies, it will be examined as a potentially useful adjunct measure to use in these studies. Thus one more analysis method will be described which will be utilized with RT:

### Bootstrapped Analysis of Reaction Time (BART)

The bootstrapped analysis of reaction time (BART) uses identical methodology to BAD with the exception that instead of brainwaves, only reaction times are considered. It has been shown that reaction time may be a useful deception detector in that reaction time to guilty probes is longer than reaction time to irrelevant stimuli (Seymour et al., 2000). In order to pose and answer this question *within individuals*, BART randomly samples, with replacement, average reaction times for the probes and irrelevants and subtracts the irrelevant average from the probe average. 100 iterations of the above process yield a distribution of 100 differences between bootstrapped average reaction time to the probe and irrelevant stimuli. If the value of zero is more than 1.65 SDs (95% confidence) below the mean of the difference distribution, then the subject is considered guilty.

## COUNTERMEASURE STUDY 1

### INTRODUCTION

This study was directed at developing a countermeasure to the Farwell & Donchin (1991) paradigm. These authors utilized a mock crime scenario with six details selected as probes, 24 details defined as irrelevants, and six other irrelevant details were utilized as targets. All items were repeatedly presented. Responses to all probes, targets and irrelevants were separately averaged by category into separate P3 averages for the comparisons as in BC-AD noted above. It is noted that the subjects used by Farwell & Donchin were paid volunteers, including associates of the experimenters. Our presently reported study uses introductory psychology (“off the street”) subjects, more like the subjects one might find in the field in the sense of relative lack of motivation and perhaps intelligence. Farwell and Donchin’s decision to use six probes, one may surmise, is based on the implied recommendation by GKT expert, Lykken (1981), that at least six items are necessary in the GKT(CIT) to have classification accuracies >90% for both innocent and guilty subjects. The idea is that a guilty subject should be shown to respond to each of a plurality (say 4/6) of stimuli. In fact, by combining responses to all six probes into one average, as in Farwell & Donchin (1991), it is possible that a guilty conclusion can be falsely obtained because the subject is responding to only one or two of the six items. This is particularly so using their BC-AD method: The two responded-to items generate a clear P300, the other four probes do not and produce a straight line average. Averaging all six yields a small P300, but the correlation method scales amplitude, so that the probe and target averages will show a high cross-correlation. We shall return to this point later.

### METHODS

#### Subjects

The subjects in this experiment were randomly selected undergraduates at Northwestern participating in order to fulfill a course requirement. All had normal or corrected vision. Subjects were randomly assigned to one of three groups. There were a total of 33 subjects (11 per group) after six subjects were lost due to high blink rate (N=4) or failure to follow instructions (i.e., failure to press yes to the targets >10% of the time; n=2).

#### Data Acquisition

EEG was recorded with silver electrodes attached to sites Fz, Cz, and Pz. The scalp electrodes were referenced to linked mastoids. EOG was recorded with silver electrodes above

and below the right eye. They were placed intentionally diagonally so they would pick up both vertical and horizontal eye movements as verified in pilot study. The artifact rejection criterion was 80  $\mu$ V. The EEG electrodes were referentially recorded but the EOG electrodes were differentially amplified. The forehead was grounded. Signals were passed through Grass P511K amplifiers with a 30 Hz low pass filter setting, and a high pass filters set (3db) at .3 Hz. Amplifier output was passed to a 12-bit Keithly Metrabyte A/D converter sampling at 125 Hz. For all analyses and displays, single sweeps and averages were digitally filtered off-line to remove higher frequencies; 3db point = 4.23 Hz. P300 was measured in two ways: 1) Base-peak method(b-p): The algorithm searches within a window from 400 to 900 ms for the maximally positive segment average of 104 ms. The pre-stimulus 104 ms average is also obtained and subtracted from the maximum positivity to define the b-p measure. The midpoint of the maximum positivity defines P300 latency. 2) Peak-Peak (p-p) method: After the algorithm finds the maximum positivity, it searches from P300 latency to 2000 ms for the maximum 104 ms negativity. The difference between the maximum positivity and negativity defines the p-p measure.. We have clearly and repeated shown that using the BAD method, p-p is a better index than b-p for diagnosis of guilt vs. innocence in deception detection (e.g., Soskins et al., 2001).

### Experimental Design and Procedures

This study was a replication and extension of the study by Farwell and Donchin (1991) with nearly identical procedures. Subjects were trained and then performed one of two different mock crime scenarios while unaware of the existence of the other. Thus each subject could be tested on both the guilty scenario (the one performed by the subject) and the innocent scenario (the one not performed by the subject). With each scenario were associated six specific details (later to be used as the probes), knowledge of which indicated the participation of the subject in that scenario. One scenario involved stealing a ring with a name tag out of a desk in the laboratory. Probes included the item of jewelry stolen, the color of the paper lining the drawer, the item of furniture containing the ring, the name of the ring's owner, etc. The other scenario involved removing an official university grade list for a certain psychology course taught by a specific instructor mounted on a blue colored construction paper, posted on a wall in a certain room. Probes included what was stolen, the color of the mounting paper, the name of the course, etc.

In order to insure awareness of the relevant details, the training of a subject involved several repetitions of the instructions followed by tests that subjects passed before beginning the lie test. Following successful completion of the instructional knowledge test and performance of the mock crime, subjects underwent an ERP-based CIT (GKT) for knowledge of each of the two scenarios, both guilty and innocent. The order of testing was counterbalanced across subjects.

During the ERP based CIT, stimuli consisting of single words were presented visually on a monitor in front of the subject for the duration of 300 ms. The inter-stimulus interval was 3048 ms, of which 2048 ms were used to record the ERP. (These timing parameters were chosen as they were used in the most recent embodiment of the Farwell & Donchin, 1991 paradigm as described by Farwell & Smith, 2001.) Subjects were instructed to press one of two buttons in response to each stimulus. In response to stimuli designated as targets, subjects were instructed to press a different button than in response to all other stimuli. The subjects were not instructed regarding the fact that some of the non-target stimuli were probes while others were irrelevants. For each of the two scenarios, subjects were tested with at least 180 trials. On each trial, subjects were presented with one word that was a target, a probe, or an irrelevant, depending on the trial.

These were presented in a Bernoulli sequence such that there were no cases in which one word was repeated twice in a row. The set of targets contained six words each of which was repeated five times in each block so that the total number of target trials was at least 30 or 17% of the total number of trials. For each target, there were four similar irrelevants presented, for a total of 24 unique irrelevants. Each of the irrelevants was repeated five times for a total of 120 or 66% total of irrelevant word presentations. There were also six probes for each scenario, each of which was repeated five times per block yielding at least 30 or 17% of the total number of trials. It should be noted that data rejected due to ocular artifact were not used and more trials were collected, yielding perhaps more than a total of 180 trials experienced.

Prior to each block, subjects were shown a list of the target stimuli for that block and asked to repeat and memorize the list. They were then instructed to press one button in response to those stimuli and another button in response to all other stimuli during the task as soon as the stimuli appeared on the screen.

Subjects were randomly assigned either to the *innocent* group, countermeasure(*CM*) group or the no countermeasure, simple guilty group (*guilty*). The only difference between the latter two groups is that those in the countermeasure group were informed about a countermeasure. They were instructed to perform covert acts in response to the irrelevant stimuli. More specifically, subjects in the countermeasure group were instructed to perform the acts in response to irrelevant stimuli based on stimulus groups or categories, rather than to individual stimuli. For example, for the name-of-operation category, animal names were used. For the probe “Cow” and target “Pig”, the irrelevants were: Horse, Goat, Sheep, and Mule. In this case, the same covert countermeasure behavior was performed every time the subject saw any of these irrelevants, as they are all in the same irrelevant group or animal name category. The covert acts were 1) pressing the left forefinger imperceptibly greater on the leg where the left hand rested, 2) Imperceptibly wiggling the big toe in the right shoe, 3) imagining the experimenter slapping the subject in the face. There was no act for the fourth irrelevant stimulus, which would thus become an oddball in its own right as the only irrelevant stimulus *not* requiring a particular covert response.

#### Analyses

In order to determine success rate of the countermeasure, BAD with base-peak and peak-peak P300 and BC-AD were performed. The use of these analysis methods also allowed comparisons with regard to efficacy and resistance to countermeasures. It is noted that since Farwell & Donchin (1991) utilized a 90% confidence interval with the BC-AD method they introduced, we also utilize a 90 confidence level for BAD analyses in this study for purposes of comparison of methods. Additionally, for the first time, analysis with BART was performed. The innocent group served to measure false positive rates for each analysis method. For each analysis method, one is considered to have appeared innocent if one’s behavioral data show that one paid attention to the stimuli, and if the analysis method did not yield a guilty result. This was the case with each analysis method employed. We can therefore objectively evaluate each analysis method in terms of hit rate, false positive rate, and resistance to countermeasures.

EXPECTATIONS:

1. It was expected, based on previous work, that both the BAD and BC-AD methods of analysis would detect at least 80% of the subjects in the simple(no CM) guilty group, a significantly reduced proportion of subjects in the CM group, and 0-10% of the innocent subjects.
2. It was expected that reaction time (RT) to all stimuli, but particularly to the irrelevant stimuli, would be elevated in the CM group since only these subjects had to decide which CM, if any, to execute on a particular trial.

RESULTS

Behavioral: Subjects followed instructions as indicated by the fact that proportions of erroneous responses to the three categories of stimuli were well under 10% seen in the following table of error rates to the three stimulus types for the three groups:

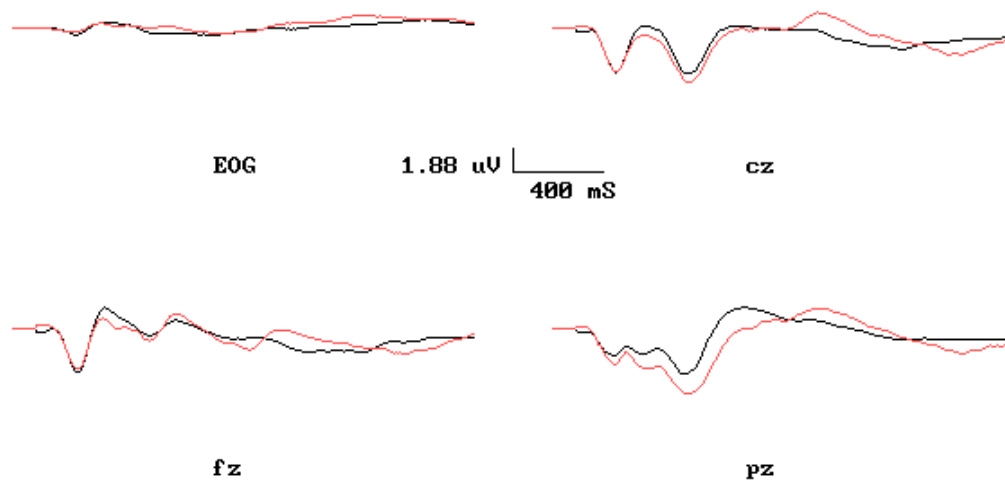
**TABLE 1:** Error rates

	Targets	Probes	Relevants
Guilty group	4%	1.5%	4.2%
Innocent group	5.2%	0.1%	0.1%
Countermeasure group	2.6%	0.5%	6.6%

Reaction Time data will be considered later.

ERPs; qualitative: The figure below shows grand averages in the guilty group for superimposed probe and irrelevant responses. In this and subsequent figures, positive is down and “Count” represents the numbers of sweeps per average. R is for probes, W for Irrelevants, and TR for targets. (The filenames at the upper left are of no concern to the reader.)

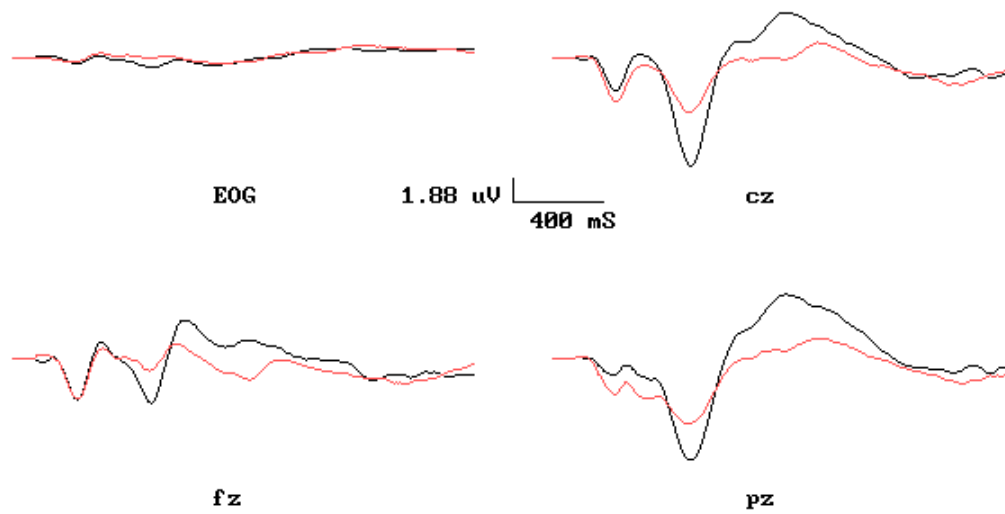
File FAR10006.GA6 FAR10006.021 Tag W R Count 951 261



**Figure 1:** Grand average ERPs in the guilty group; R = probe, W = irrelevant at 4 sites as indicated. Positive is down in all ERP figures.

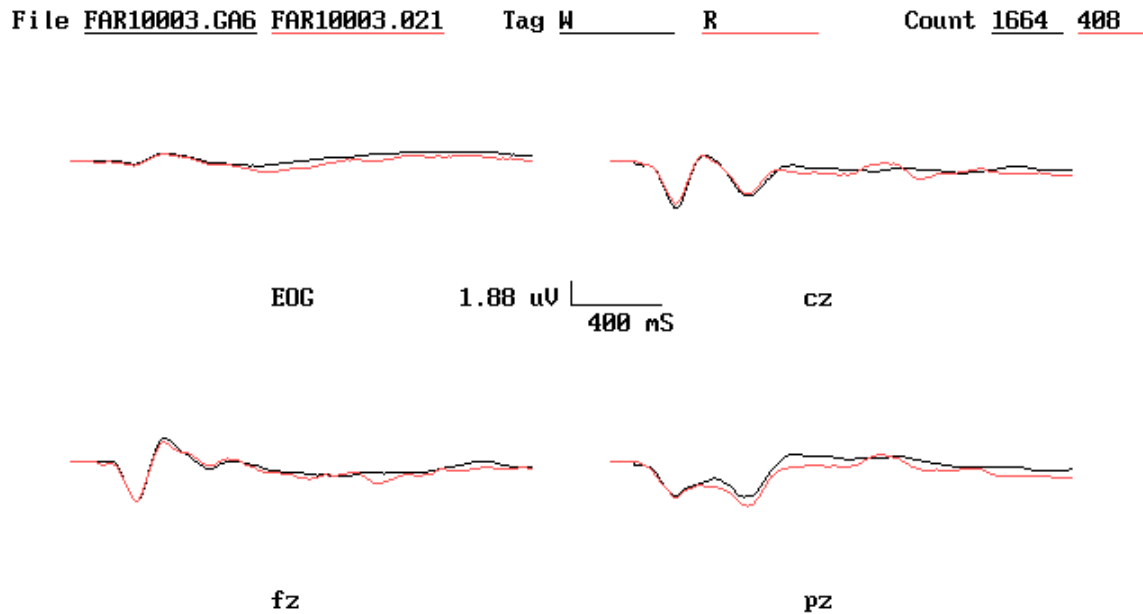
It is as expected that a larger P3 response is seen to the probe than to the irrelevant at Pz. The following figure shows superimposed TR and R responses in the guilty group, and although the TR is larger, the morphology of R and TR are similar.

File FAR10006.GA6 FAR10006.021 Tag TR R Count 237 261



**Figure 2:** Superimposed grand averaged R and TR responses, guilty group.

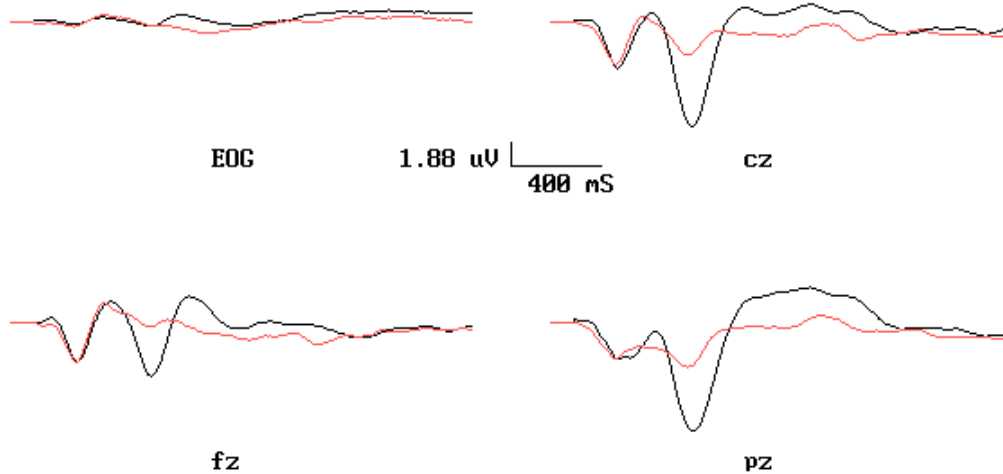
The following figure shows superimposed ERPs to W and R in the innocent group, and it is clear that there is little difference between the P300s as expected since for the innocent, the probe is just another irrelevant.



**Figure 3:** Innocent group. Superimposed grand averages to probes and irrelevants.

In the next figure, the targets and probes in the innocent group are shown, and the target P3s tower over the irrelevants as expected. This is the prototypical innocent picture.

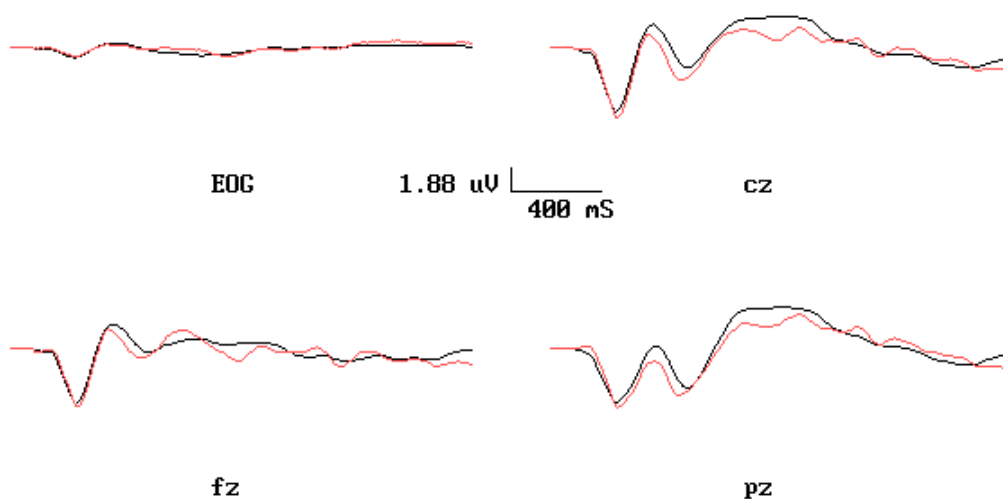
File FAR10003.GA6 FAR10003.021 Tag IR R Count 477 408



**Figure 4:** Superimposed grand averages to targets and probes, innocent group.

The expected effect of the countermeasure is shown in the following figure in which probes (R) and Irrelevants (W) are virtually identical in the countermeasure (CM) group. Of course they are superimposed in the innocent group also (so the countermeasure users appear innocent), but there is more of a P300 for R and W in the countermeasure group. This is probably because in the innocent group the probe is just another irrelevant, but in the countermeasure group the probe is relevant because the subject is guilty, yet the irrelevants have also been made task-relevant by the covert responses:

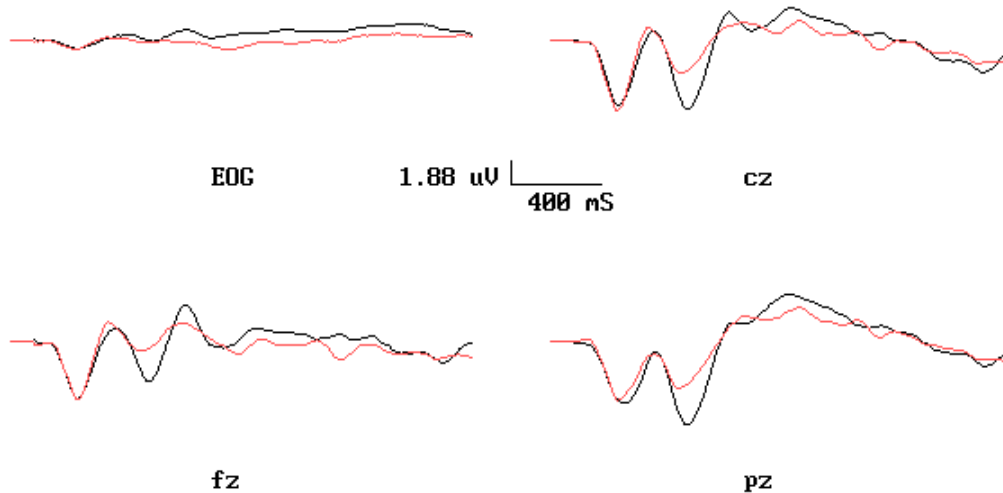
File FAR10004.GA6 FAR10004.021 Tag W R Count 1115 318



**Figure 5:** Superimposed probe and irrelevant grand averages, CM group.

Finally, the next grand average figure shows probe and target and P300s in CM group.

File FAR10004.GA6 FAR10004.021 Tag TR R Count 300 318



**Figure 6:** Target (TR) and probe [R] grand average responses in countermeasure group.

The countermeasure has produced the desired effect in that the target P3 clearly exceeds (by 2.25 uV, b-p or p-p) the probe which is about the same as the irrelevant (Figure 5). This is the innocent look. Because we will refer to this effect in the discussion, we will present one quantitative result here: A paired t-test on the probe vs target P300 yielded  $t(10) = 2.48$ ,  $p < .04$  b-p, but  $t(10) = 1.66$ ,  $p = .12$  p-p. It is further noted that in b-p P300, 10/11 target responses were substantially larger than probe responses. For p-p, the proportion was 9/11.

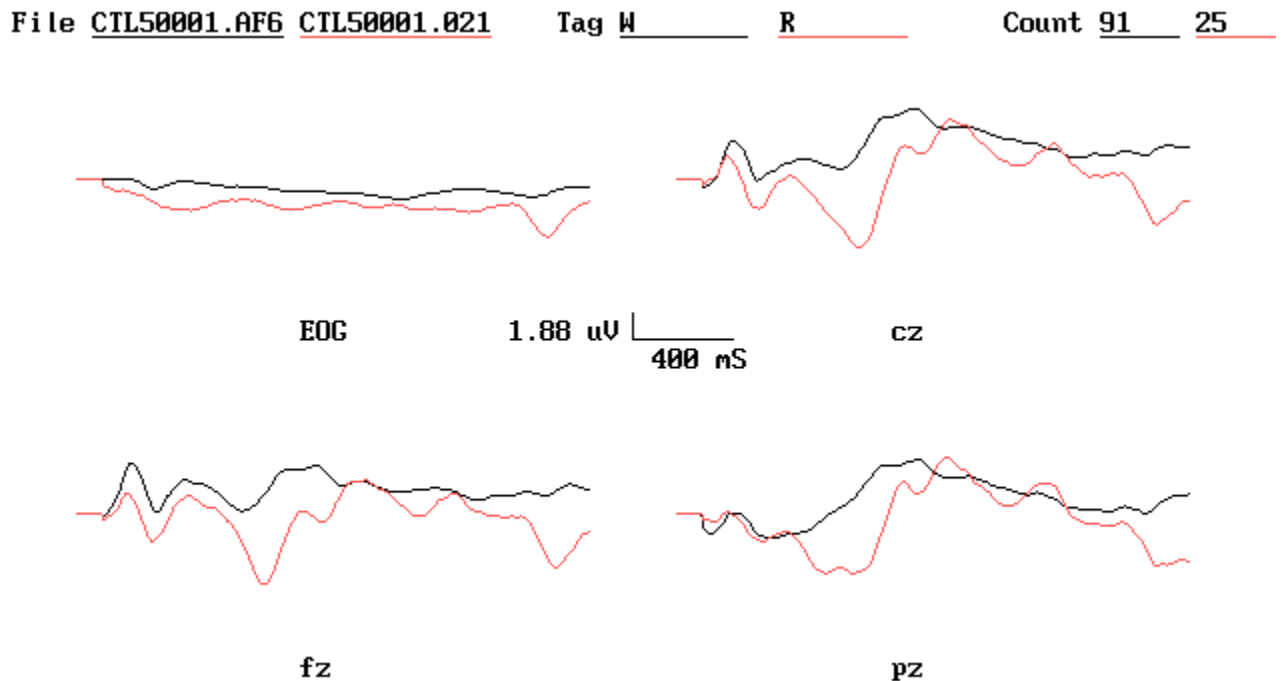
ERPs, Quantitative analysis: The table below gives the proportions of guilty calls as a function of group and analysis method (BAD and BC-AD). Also presented are results with BART, an analysis of the differences in reaction times, probe minus irrelevant, as it is expected that the reaction time to the irrelevant stimuli should increase due to performance of the covert acts in the countermeasure (CM) group.

**TABLE 2:** Outcomes of first CM study.

Group	BAD(b-p)	BAD(p-p)	BC-AD	BART
Guilty	8/11 (73%)	9/11 (82%)	6/11 (54%)	10/11(91%)
Innocent	1/11 (9%)	1/11 (9%)	0/11 (0%)	1/11 (9%)
CM	2/11 (18%)	2/11 (18%)	6/11 (54%)	5/11 (45%)

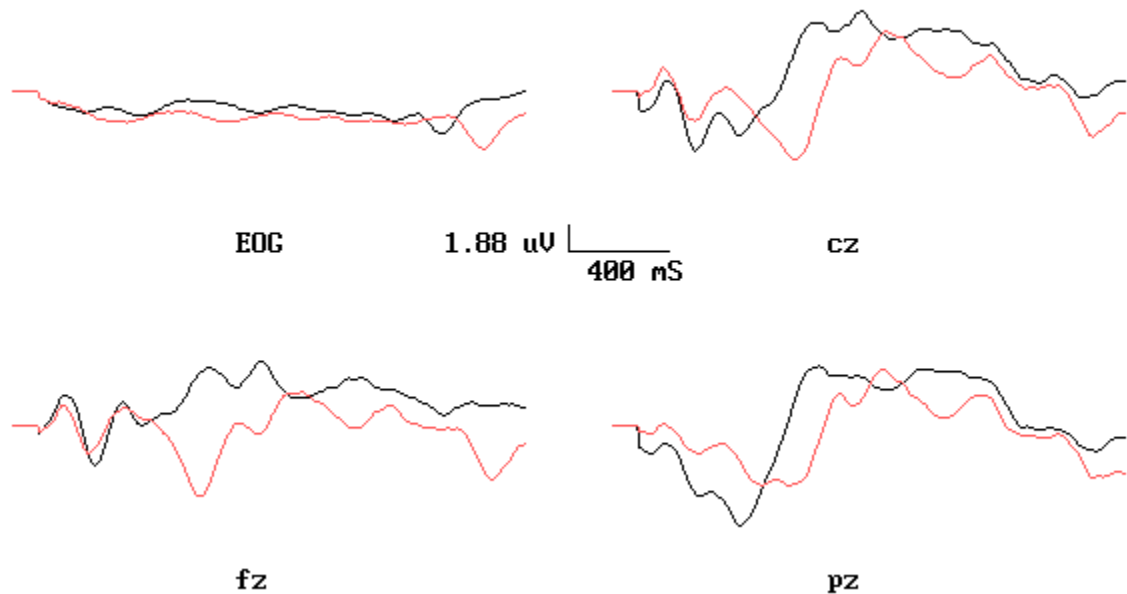
What emerges first from these data is that most of the guilty group (82%) are detected with BAD on p-p values, and as usual (Soskins et al., 2001), p-p outperforms b-p. The false alarm rate using BAD on p-p data is a low 9%. Thus the manipulations appear to be working, lending credibility to the effect of the countermeasure which (with BAD, p-p) reduces the 82% hit rate in guilty subjects to 18% in guilty subjects using the countermeasure. ( $p=.08$ , Fisher exact test). Secondly, it is clear that the BC-AD performs poorly (near chance levels at 54%) even in these guilty subjects. BAD (p-p) outperformed BC-AD at  $Z=2.45$ ,  $p<.05$  on McNemar's test of differences between correlated proportions.

We believe the poor performance of BC-AD here (vs 87.5 % hit rate in Farwell & Donchin, 1991) is attributable to the greater P3 latency variance one might expect to see in the unmotivated naïve subjects run in the present study, and the motivated, paid subjects of Farwell & Donchin, (1991). In particular, differences in latency between Target and Probe stimuli could lead to out of phase ERPs; BC-AD, which looks at the simple cross correlation of Probe and Target could find low cross correlation coefficients between such out of phase responses. This situation, which leads BC-AD to a miss decision, is illustrated in the next 4 figures. Figure 6a shows superimposed probe (R) and irrelevant(W) responses in a guilty single subject of this experiment; he is clearly guilty and that is the outcome of the BAD test. However, Figure 6b shows the superimposed Target (TR) and R responses as well out of phase, and thus BC-AD failed to detect this subject.



**Figure 6a.** Average ERPs from a single guilty subject. R clearly > W at all 3 sites.

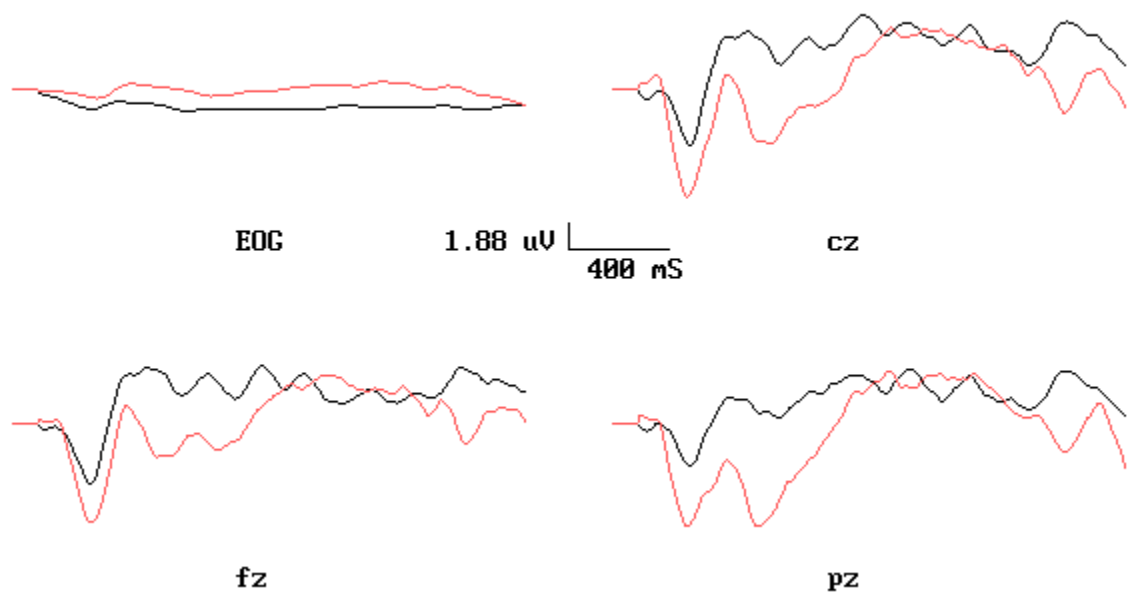
File CTL50001.AF6 CTL50001.021 Tag TR R Count 22 25



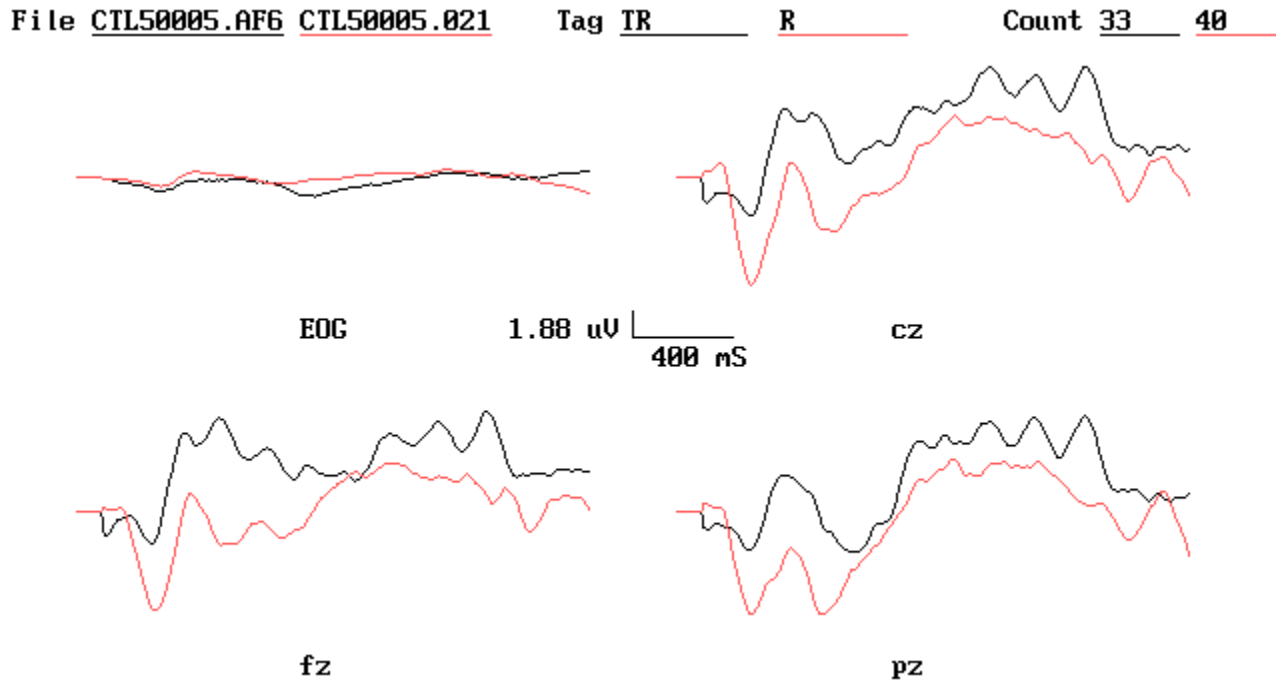
**Figure 6b:** Superimposed TR and R average responses from the same guilty subject as in previous figure; note striking P300 phase shifts evident at Cz and Pz.

Figures 6c and 6d, below, are comparable to 6a and 6b, respectively, but are from another subject.

File CTL50005.AF6 CTL50005.021 Tag W R Count 111 40



**Figure 6c:** Same as 6a, different subject.

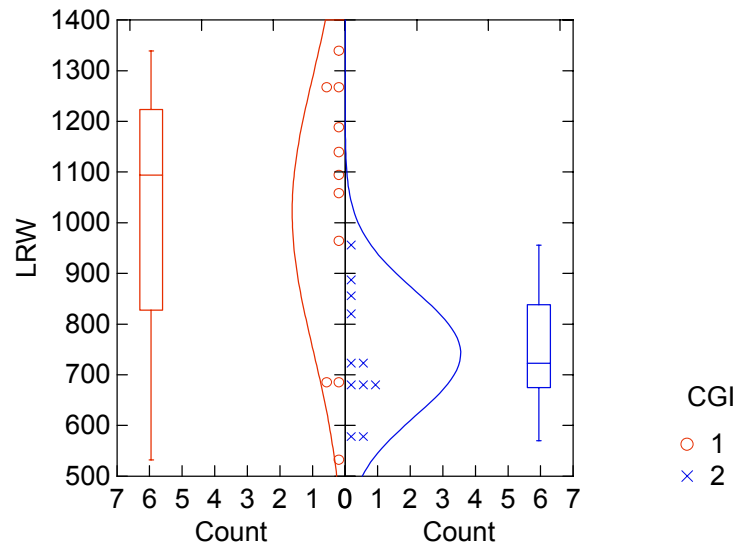


**Figure 6d:** Same as 6b, same subject as 6c.

Indeed, even the respectable 82% hit rate seen here with BAD (p-p) in guilty subjects is about 5-10% lower than we generally report using the single probe paradigms described in the next study. It could also be the case that the 6 probe paradigm is more complex than the 1 probe paradigm, producing more task demand which depresses P300. It is incidentally noted that Allen & Iacono (1997) reported that the BC-AD method was slightly more accurate than the BAD method. This is not really contradictory as Allen & Iacono were using paid subjects and their BAD analysis was on b-p amplitudes, which we know to be up to 30% less accurate than p-p amplitudes in ERP-based CITs (Soskins et al., 2001).

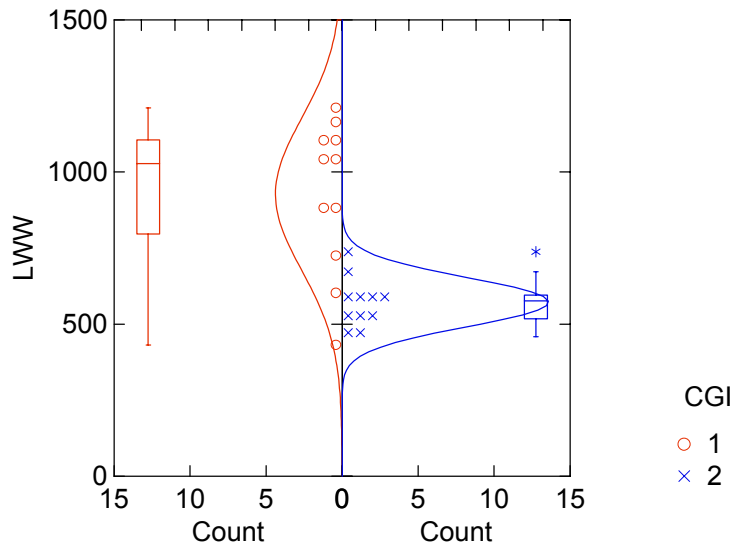
To get some evidence on these matters, we ran a separate (6 probe) study exactly like the present guilty group, except that the 14 subjects were advanced, motivated sophisticated subjects in an upper level lab course. In this study, both BC-AD and BAD (at a 90% level of confidence) detected 10/11 (91%) subjects. (Three subjects were dropped for having target error rates >10%.)

Finally, regarding Table 2, it is noted that the reaction time data indicate that although RT correctly classifies 91% of the guilty subjects, that figure is halved to 45% when the countermeasure is used. These data are based on the R-W (probe-target) difference. It might be suggested that the expectedly increased *absolute* value of RT for either probe or irrelevant in the CM group could alert an examiner that countermeasures are being used. As the following figures will make clear, however, although this expectation is borne out by group analysis, it is *not* in individual analysis:



**Figure 7:** These are reaction times and fitted distributions to the dishonest probe no” responses (LRW) for both countermeasure (left) and guilty (right) groups.

The variance is clearly greater in the CM group, as is the mean RT ( $t=3.1$ ,  $df=20$ ,  $p<.008$ ), however it is evident that at least 4 members of the CM group fall into the distribution of the guilty group, and all 4 of these subjects were successful in defeating the test (i.e., they were undetected by BAD, p-p). Indeed the three slowest members of the countermeasure group--in the RT sense--failed to beat the test; (slow responders seem to be slow countermeasure learners). The following figure shows the similar pattern of results with respect to the irrelevant stimulus:



**Figure 8:** RTs to irrelevant/'no' stimuli (LWW) in countermeasure(left) and guilty (right) groups.

Again, the difference between means is significant ( $t=4.5, df=20, p<.001$ ), however, it is clear that for any given individual undetected by BAD, one can never be certain from RT whether or not a countermeasure (CM) is being used. One would like to think that certain very high RT values are typical of CM-users, however, as we have seen, the slowest responders tend not to escape detection with the P300 test. Again, it is emphasized that these data are from unmotivated, unsophisticated subjects. It is likely that well-trained, highly motivated extremists would be practiced and fast, even with their CMs. Not shown are the RT distributions comparing innocent and CM groups. The figures would look extremely similar to Figs. 7 and 8 just shown, indicating that the fastest CM users who beat the test look just like innocent subjects in RT.

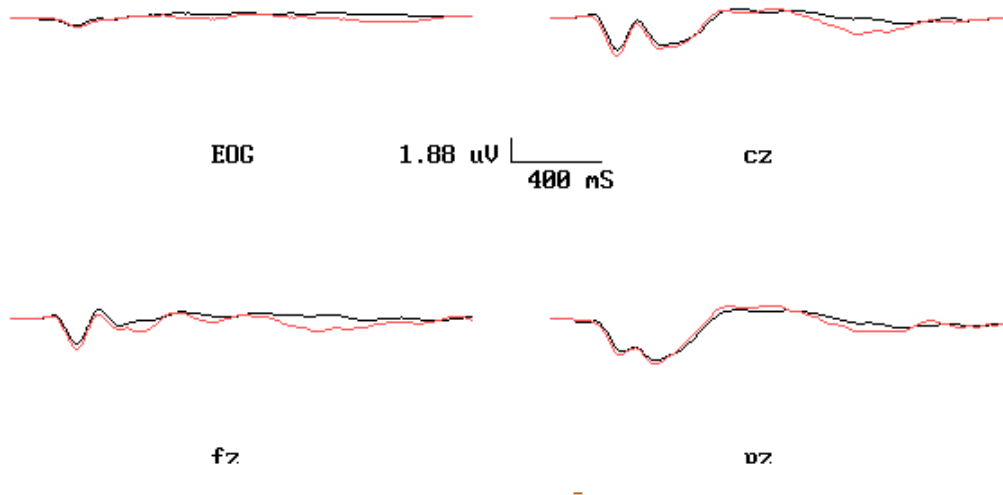
## DISCUSSION: INTRODUCTION TO COUNTERMEASURE STUDY 2.

The previous study showed that the 6-probe paradigm of Farwell & Donchin (1991) may be significantly impacted by the countermeasure (CM) of making irrelevants secretly relevant. The result was that the probe and irrelevant responses became largely indistinguishable in the guilty subject employing a CM successfully (as in Figure5, a grand average figure which well represented most individuals). Both stimuli evoked reduced P300 responses of about the same small size. One could argue that an investigator could become suspicious in such a case because theoretically, one would not expect *any* P300 response to irrelevant stimuli. However, the fact of the subjects' cooperation would be supported by the accurate response rates (>90%) to the target stimuli. Also, it turned out that the target responses in the CM group (Figure6) were larger than the probe responses, which would make it very difficult to press the case that the subject was guilty, but using a CM. The large target response would indicate a normal P300 to a sole

oddball and a cooperative subject. One would have to perhaps conclude that the subject was aberrant in the sense of having a small but distinct P300 to irrelevants and probes, but one could not conclude guilt. Clearly, however, an ideal CM would make the subject's responses look like those in Figs. 3 and 4 above, the responses of an innocent subject, in which the target response towers over the probe response which contains no or a very small P300 response, comparable to the irrelevant response. We will refer below to such a pattern as a *classical defeat* of the test.

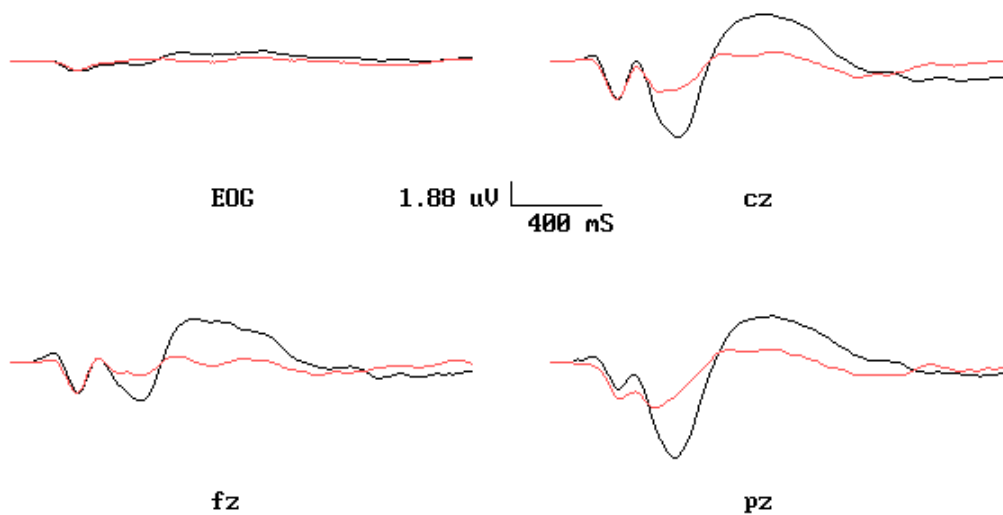
Relevant to this point, we had done a pilot study (with 10 advanced subjects in a lab course) in 2001 in which we utilized the 1-probe paradigm on which we have often reported (e.g., Soskins et al., 2001). In it, the subject sees on every trial a sample from a random series of dates, one of which is his birth date. He is instructed to deny recognition on an unpatterned half of the trials (so as to simulate instructed head injury symptoms). One irrelevant date is designated a target date as above. In an initial simple guilty run with 9 total stimuli, BAD and BC-AD both detected >88% of the subjects in two runs each (stimulus duration was either 300 or 2048 ms). In a subsequent week with an unsuccessful countermeasure (count backwards by 7s until a target is detected, then resume count), the two analysis methods detected >95% of the subjects in two runs each. In the following week, we imposed 2 changes: 1) We reduced the number of stimuli to 6, 2) we instructed the subjects to perform the same CMs as in EXPERIMENT 1 above to the first three encountered irrelevants. (Again, multiple runs were used.) Now the hit rate was 42 % with BC-AD and 37% with BAD. Clearly, the CM was effective, however this time, target, probe, and irrelevant stimuli all produced about the same sized smaller P300 in the grand average. We were satisfied that the CM worked, though not ideally as would be the case if the results looked like those in Figs. 3 and 4 above. Finally, we knew we needed an appropriate control guilty group with 6 total stimuli--we could not use the results of the first 2 weeks with 9 stimuli--so we instructed the subjects to repeat the 1-probe, 6-stimulus paradigm, but to stop using the CM. They appeared to follow instructions as indicated by the target response accuracy, but the hit rates remained low at 29% (BC-AD) and 58% (BAD). It was as if they could not stop using the CM, although all reported no effort to use it, and their RT data indicated no hint of a CM, use of which was apparent when they did use the CM the previous week (We will show such RT data below.) The really interesting finding, however, was that in 14 of 22 runs (on 7 subjects), we achieved the *classical defeat* pattern, as shown in the following 2 figures:

File THR10008.GA6 THR10008.022 Tag W R Count 1974 511



**Figure 9:** Grand average probe(R) and irrelevant (W) virtually identical small responses from 14 runs of 7 guilty subjects in pilot study instructed to not use the countermeasure learned the previous week.

File THR10008.GA6 THR10008.022 Tag TR R Count 501 511



**Figure10:** Similar to Figure 9, except probe [R] and target [TR] are superimposed and TR towers over R, a classical defeat.

It would appear that somehow, the experience of learning the countermeasure --just once-- appears to release the response to targets, while keeping probe responses suppressed, in subsequent tests, and thereby produces a classical defeat of the P300-based CIT. The mechanism of this effect is of theoretical interest to the ERP community, and will be pursued elsewhere. The study to be described now was a more formal and systematic attempt to replicate these findings of the pilot study. Additionally, we felt that the 1-probe paradigm is essentially simpler to use as a research tool with respect to development of a good countermeasure. Moreover, we will argue below that a series of 1 probe runs with a different probe each time is more interpretable than the 6-probe paradigm, and is also more susceptible to CM detection via the use of reaction time.

## METHODS

Subjects. The subjects in the experimental group were 14 members (5 female) of a junior-senior level advanced class in psychophysiology. All had taken and received B+ to A grades in two previous quarters of a neurobiology class. All had normal or corrected vision. A control group (no CM) of 10 paid volunteers (5 graduate students, 5 females) was also run.

Procedure: The experimental group was run through the 1-stimulus birthday paradigm described above in three successive weeks. In the first week, they were completely naïve about the countermeasure, and were told that the first experiment was simply to demonstrate the ability of the P300-based CIT to detect behaviorally denied autobiographical information. In the second week, they were instructed in the CM. They were specifically told to execute the covert finger press upon encountering the first non-target, non-probe, i.e., irrelevant stimulus; the covert toe wiggle upon encountering the second irrelevant, and the mental visualization of being slapped by the instructor upon encountering the third irrelevant. They were told explicitly to do nothing upon seeing the fourth irrelevant, which would take care of itself by being the only irrelevant stimulus requiring no response—an oddball in that sense. In the third week, the subjects were told to perform without the countermeasure, as they had done the first week. The control group was simply run as in the first week for the experimental subjects, and they were run that way for three successive weeks. We told them, truthfully, that we were interested in possible changes in the patterns of responses over a three week period. The timing parameters of stimulus presentation and duration were just as in the first experiment. For all subjects, each run consisted of a minimum of 180 trials with each of the 4 irrelevant, 1 probe, and 1 target stimuli repeated a minimum of 30 times, yielding average ERPs of at least 30 sweeps each.

EEG and data analysis methods: These were exactly the same as in the previous experiment.

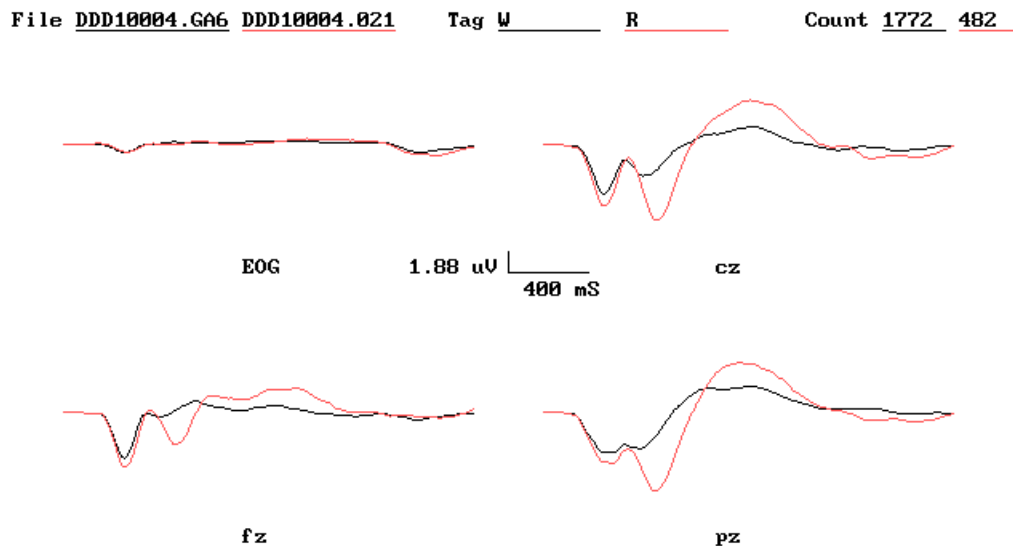
## EXPECTATIONS:

1. It was expected that at least 80% of the subjects would be detected by both BAD and BC-AD in the first week.
2. It was expected that the hit rate would be significantly reduced during the second week when the CM was employed. All three stimuli were expected to produce a P300, with the probe and irrelevant P300s being indistinguishable.. It was expected that RTs to all stimuli would be raised, especially to the irrelevants, during this second week as subjects decided which CM to execute.
3. It was expected that ERPs and RTs would return to normal in the third week.

## RESULTS:

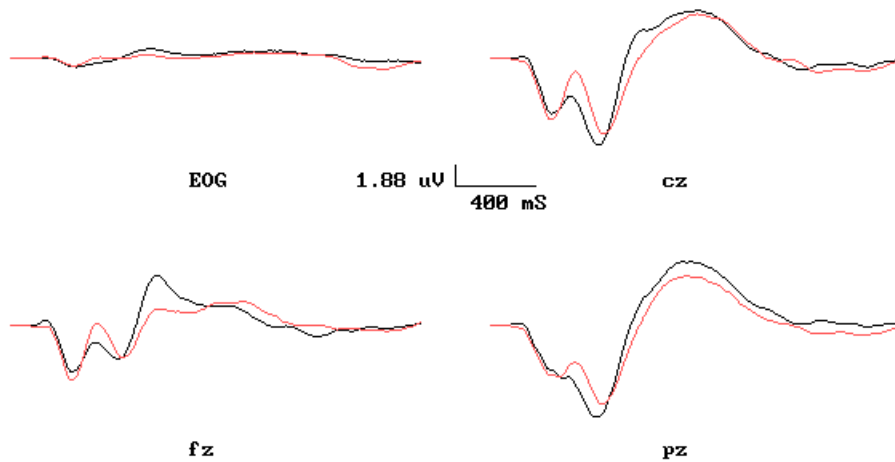
Behavioral. RT data will be presented later. That the subjects followed instructions is evidenced by the fact that only one subject in the first week had a target error (a “no” response) rate > 10%. (His data were not used.). The average target error rates for weeks 1-3 on all subjects were 6.8%, 2.0%, and 6.1%, respectively; these differences failed to reach significance in a 1 x 3 ANOVA. Errors to probes would be *truthful* (“yes”) responses; the proportion of these were low also in weeks 1-3, respectively: 0.8%, 1.3%, and 0.5%, respectively (no significant difference). Errors to irrelevant would be “yes” responses also. The rates in weeks 1-3 were 0.3%, 0.1%, and 0.1%, respectively (no significant difference.).

ERPs: Qualitative: Figs. 11 and 12, below, are the grand averages from the run of the first week in experimental subjects:



**Figure 11:** These are the superimposed probe[R] and irrelevant [W] grand averages from the first week of the 1-stimulus experiment. Labeling is as in ERP figures above.

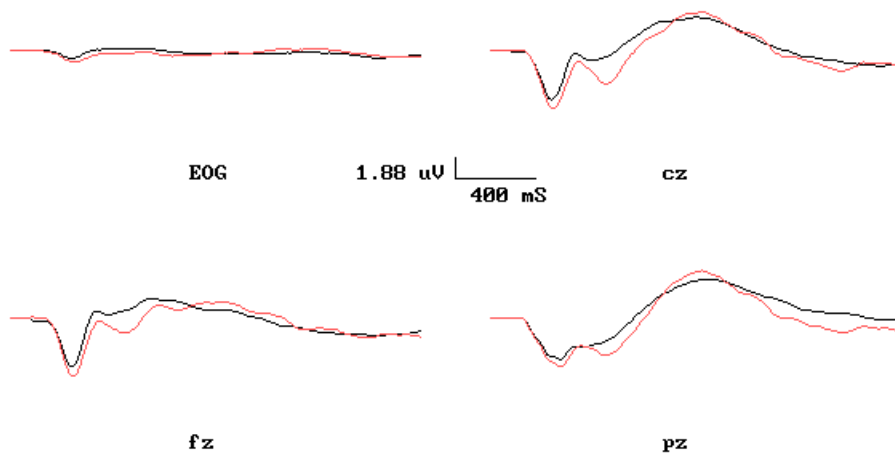
File DDD10004.GA6 DDD10004.021 Tag TR R Count 482 482



**Figure 12:** This is the same as Figure 11 except probe[R] and target[TR] grand averages are superimposed.

This is the expected and usual result of running the 1 stimulus paradigm: the P300s of R and TR are similarly large (Figure 12) and tower over the response to W (Figure 11). The next two figures show the grand average ERPs for the second week of the experimental run, i.e., when the CM was in effect:

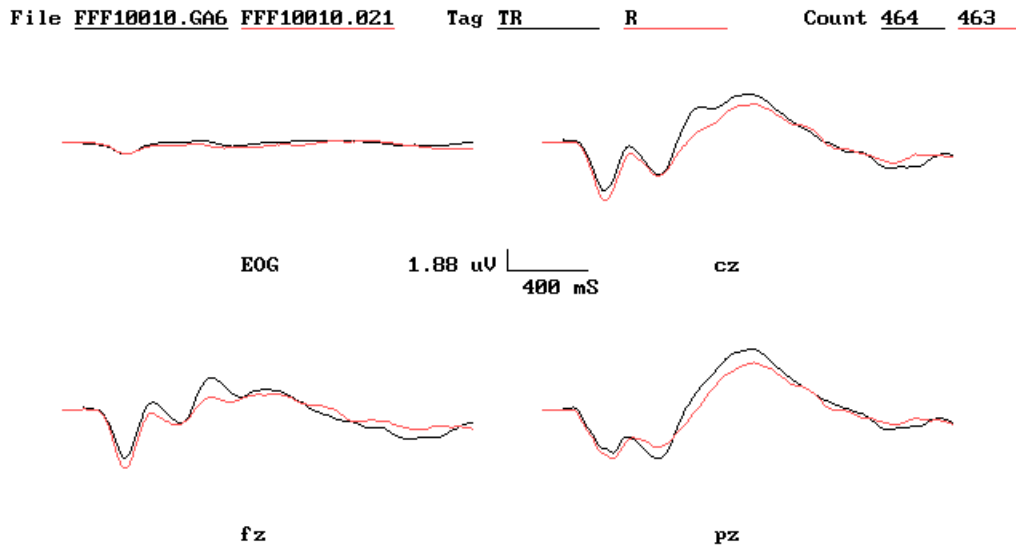
File FFF10010.GA6 FFF10010.021 Tag W R Count 1771 463



**Figure 13.** Superimposed grand averages to R and W during the explicit use of the countermeasure, experimental group, week 2.

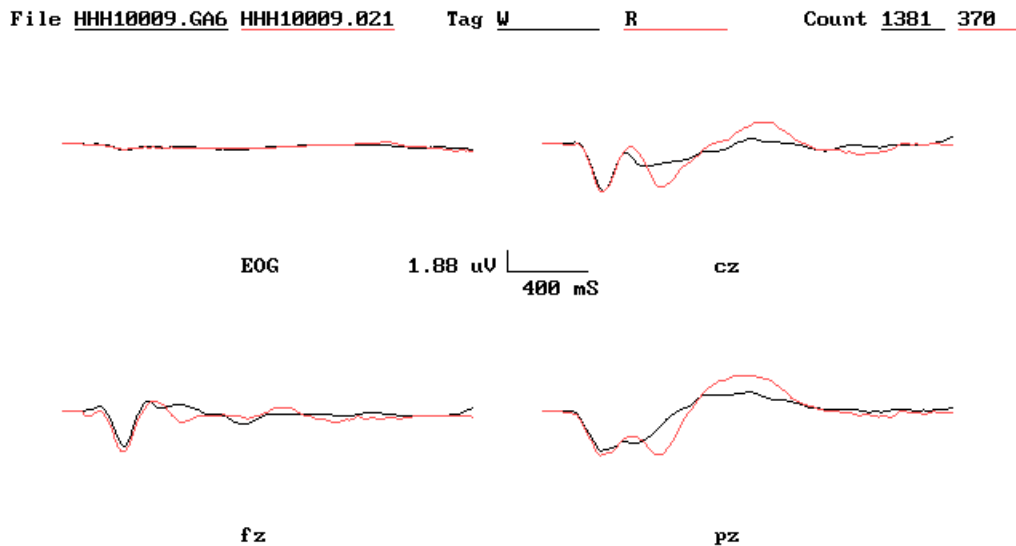
The P300s, particularly at Pz, the site where P300 is usually largest, are small and of similar size. The R is slightly larger than the W because not all of the subjects contributing to these averages successfully defeated the test. The following figure, showing the superimposed R and TR responses from the same session as in the previous figure, indicates that the targets were greatly reduced also, and are not much larger than the probes. That is, all 3 stimulus types

generated a small P300 because all were meaningful. The reduced size is no doubt due to the loss of unique oddball probability for R and TR.



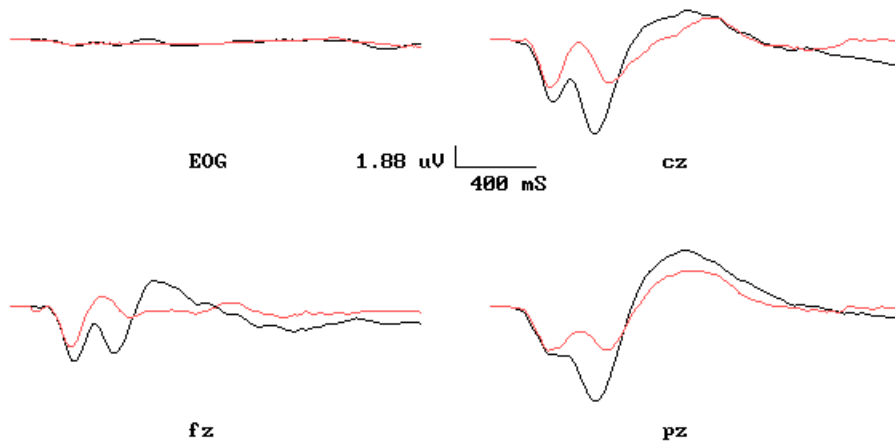
**Figure 14:** These are the superimposed grand averages of TR and R responses from same session as that represented in the previous figure. The P300s are reduced (compare with Figure 12).

The following figures show the superimposed grand averages to R and W (Figure 15) and TR and R (Figure 16) during the third run of the experimental subjects, when they were instructed to *not* use the countermeasure.



**Figure 15:** Superimposed R and W responses from third run of experimentals.

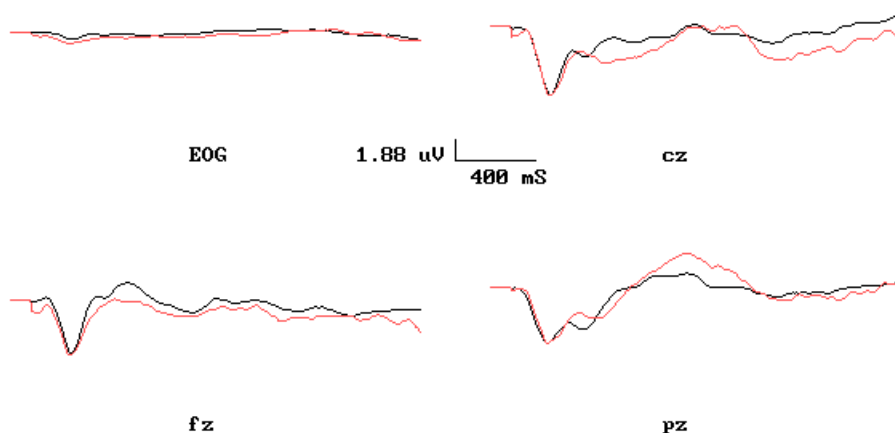
File HHH10009.GA6 HHH10009.021 Tag TR R Count 375 370



**Figure 16.** Same as Figure 15 except TR and R are superimposed.

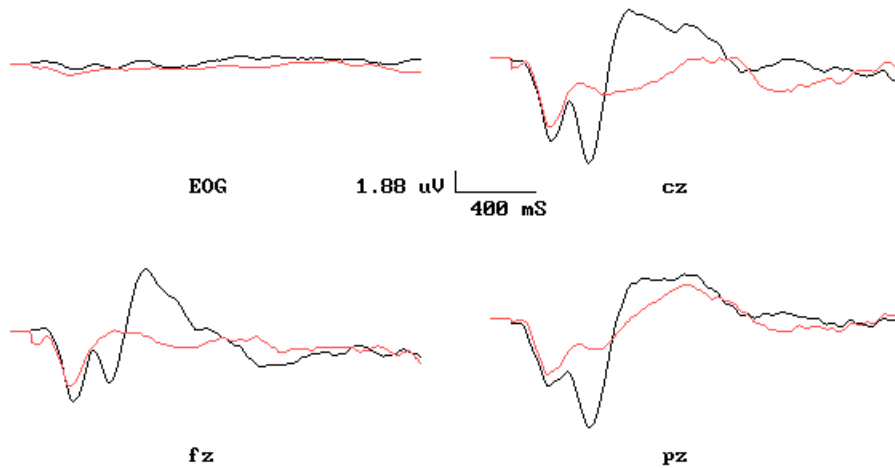
It is the case that the R response is slightly larger than the W (compare Figure 11 in the naïve subject), which is because not all subjects contributing to the grand average defeated the test in the absence of explicit use of the countermeasure. However it is clear that in the week after the explicit use of the CM, the target response is again “released” to its normally large size, relative to the probe, in all or most subjects in the third week. (A line graph will make this obvious below.). As a group, the experimental grand averages tend toward a classic defeat here, and for at least 5 subjects, the classical defeat pattern is indeed obtained., as is clear in the following two figures:

File HHH10001.GA6 HHH10001.022 Tag W R Count 496 135



**Figure 17:** Grand averages over 5 experimental subjects in week 3: The P300 for the probe is actually less positive at Pz than that to the irrelevant.

File HHH10001.GA6 HHH10001.022 Tag TR R Count 132 135



**Figure 18:** ERPs as in previous figure(same 5 subjects), except TR and R are shown and TR clearly towers over R, the classical defeat pattern.

None of the 5 cases in the preceding 2 figures were called guilty by either BC-AD or BAD (90% confidence, p-p). Not shown are the individuals contributing to these averages. Each and every one of them shows the classical defeat pattern. In another 4 subjects, although the R is somewhat larger than the W, the TR, again, towers over both R and W. As one might surmise, BAD which simply looks at R-W did successfully detect these subjects, but BC-AD did not. (Results for the control group are not shown; their response patterns for all three weeks are similar and strongly resemble those of Figs. 11 and 12 above, except that both TR and R responses declined in the third week. R responses were significantly greater than W responses in the third week ( $P < .001$ ) and 9/10 of the controls were detected by BAD.)

ERP DATA: Quantitative: The table below shows the detection rates for the experimental subjects using the bootstrap tests, BC-AD(90%), BAD (90%, p-p), and BART over the three weeks of testing:

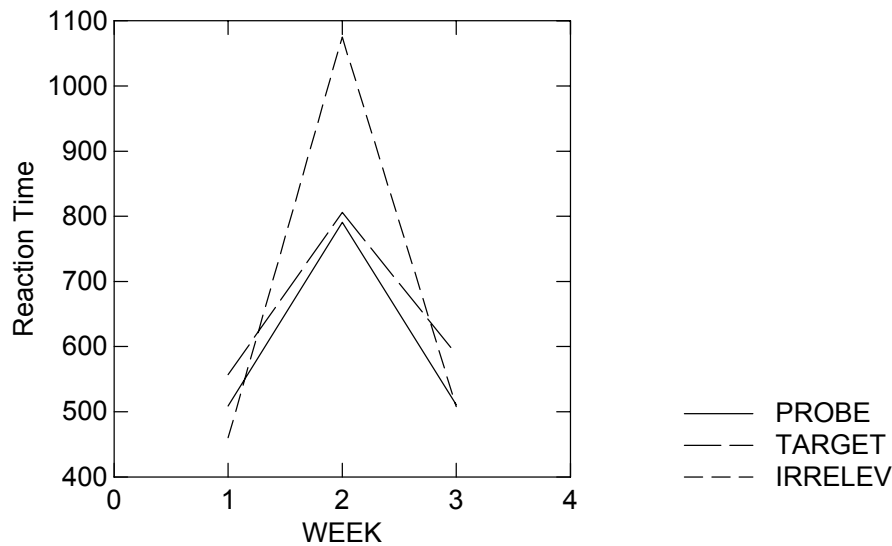
**TABLE 3:** Outcomes of second CM study

<u>WEEK</u>	<u>CONDITION</u>	<u>BAD</u>	<u>BC-AD</u>	<u>BART</u>
1	naïve, no CM	12/13(92%)	9/13(69%)	8/13(62%)
2	explicit CM	6/12(50%)	3/12(25%)	0/12(0%)
3	no CM	7/12(58%)	3/12(25%)	5/12(42%)

It is noted that in the first week, one subject (of 14) was dropped due to a target error rate >10%; in the second and third weeks, two subjects data files for the bootstrap tests were lost. What is apparent from the table are the following major findings:

- 1) BAD detects 3 of the naïve subjects (week 1) which BC-AD misses .
- 2) Using the more reliable BAD test, explicit use of the CM in week 2 drops the hit rate from 92% to 50% ( $p < .08$ , McNemar), and from 69% to 25% with BC-AD.
- 3) In the third week with the CM not used (confirmed below with RT data, and by post-experiment interviews) the hit rate is still near chance with BAD (58%), and as we saw above in the qualitative ERP data, the 5 of 12 subjects who defeat the test do so with classic defeats, appearing like innocent subjects. Indeed, the BC-AD test in the third week detects only 25% of the subjects, the same number as when the explicit countermeasure was in use. It is reasonable to speculate that more intensive practice might result in a higher proportion of such defeats. Future research on the mechanism of these classical defeats could yield more effective CM training methods.
- 4) The RT measure, BART, which looks at the difference between probe and irrelevant RT, performs poorly throughout. It is of course worthless in the second week when, as we will see, the RTs for Ws are doubled from about 500 to 1000ms (within and greater than the range of the probe RT in many cases; see Figure 19, below) as the subject must recall what CM to use following each W.

Concerning *absolute* reaction times, the following figure shows the reaction times to the three stimulus types over the course the three weeks in the experimental group:

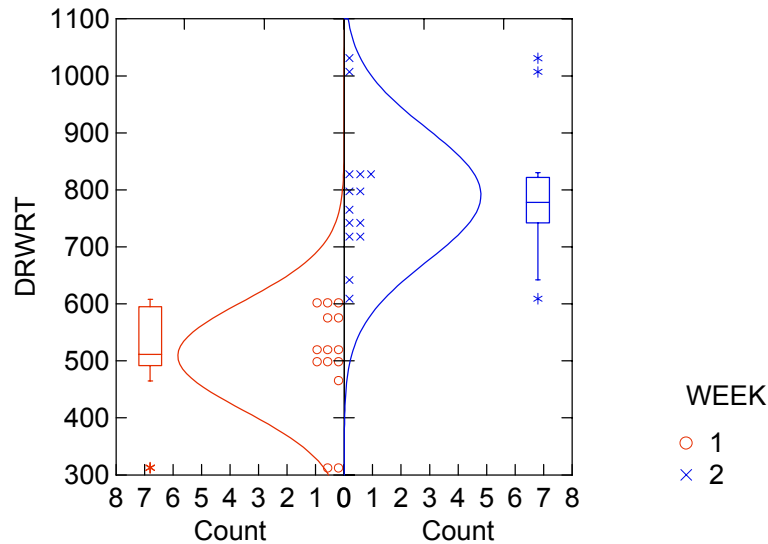


**Figure 19:** Reaction times (ms) for 3 stimulus types across 3 weeks.

Several points are implied by these data:

First, it is clear that after the dramatic increase in RTs during the explicit use of the CM, the RTs drop down in the third week to the level of the first week, providing clear evidence that the subjects followed instructions and did not use the countermeasure in the third week. A 3 x 3 repeated measures ANOVA on these scores yielded Greenhouse-Geiser corrected (GG), significant effects for all variables: Week:  $F(2,24) = 60.8, p < .001$ ; Stimulus type:  $F(2,24) = 9.47, p < .003$ ; Interaction:  $F(4, 48) = 36.5, p < .001$ . The interaction appears due to the greater increase in irrelevant than other RTs in week 2.

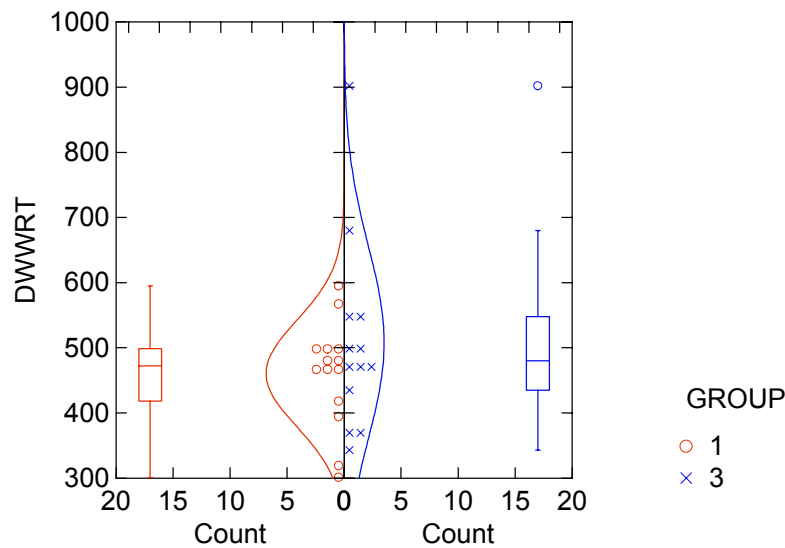
Secondly, a post-hoc ANOVA comparing RTs just in weeks 1 and 3 showed only one significant effect, that of stimulus type,  $F(2,24) = 15.7, p < .001$ , due to the expected effect of higher RTs to Target than to other stimuli, doubtless related to the need to switch response buttons for this stimulus. The point is that in the third week, when 58% of the subjects are *undetected* by the BAD test and 75% *undetected* with the BC-AD test, RT would be no help in identifying the test beater. Indeed, the 5 subjects who showed classical defeats of the ERP-based CIT, had mean RTs to all stimuli in the *lower* half of the RT distributions of all subjects. One would think, however, that RTs could at least be used to identify CM users during the explicit use of the CM. Indeed, post-hoc t-tests comparing RTs on weeks 1 and 2 yield  $t > 5, p < .001$  for all 3 stimulus types. Nevertheless, the following figure shows the RT distributions for probe stimuli in weeks 1 and 2, and it is clear that despite the significant group effect, there is still overlap. The picture is similar for the target distributions.



**Figure 20:** Reaction time distributions for probe stimuli in weeks 1 (left) and 2 (right).

*In the irrelevant RT distributions from weeks 1 to 2, however, there was in the present sample, **no** overlap, of actual reaction times from the 2 weeks, although the fitted normal distributions from both weeks included 2 subjects each from the *other* week. This suggests that an absolute value of RT to the irrelevant stimulus might be determinable which could aid in identifying explicit use of the present CM. In any particular case, it would be difficult to press this point in view of the overlapping fitted distributions. If all the RTs (to probe and target also) in a particular subject were long, that would make the identification of a CM-user all the more difficult. Indeed in a follow up study not reported further here, we showed that it was simple for these subjects to voluntarily raise RTs to *all* stimuli such that there were no significant differences between RTs to any stimuli during explicit performance of the CM. Moreover, it is possible that practice might reduce even the irrelevant RTs.*

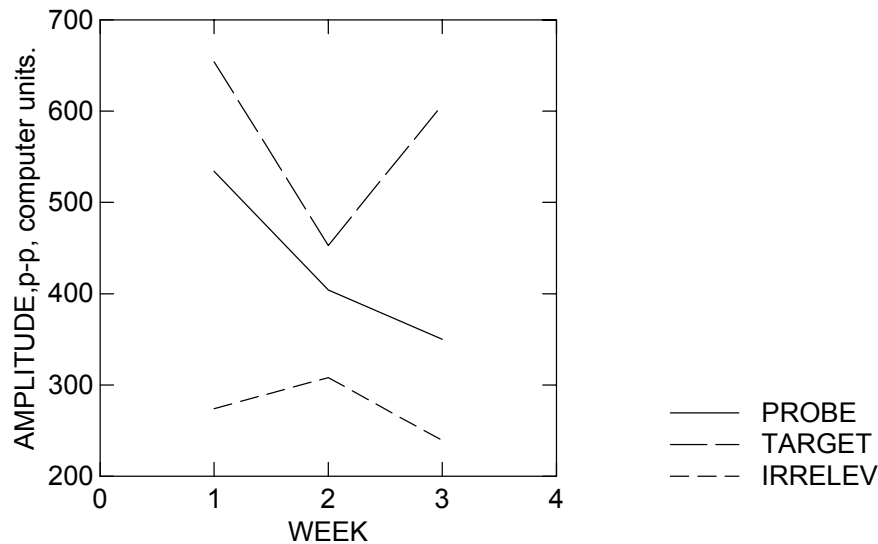
Of course one does not see a classic defeat of the ERP test in week 2 anyway, although the ERP analysis methods are largely defeated in week 2. The classic defeats of the test are seen in the third week, when the distribution to the irrelevant are as shown in the next figure, along with the distribution in the CM-naïve subject's first week:



**Figure 21:** RT distributions in weeks 1 (left) and 3 (right) to irrelevant stimuli.

The gross overlap, expectedly yields  $t(12) = 1.15$ ,  $P > .25$ . The results were very similar with the other stimuli.

The figure below plots the computer-calculated p-p means to the three stimuli during the 3 weeks for all subjects. The p-p means are shown, rather than b-p means, because both BAD and BC-AC analyses look at both the down-going P300 as well as its recovery as a negative overshoot, i.e., p-p:



**Figure 22** Computed mean p-p amplitudes, all subjects, in computer units (10 microvolts = 409.6 units) for 3 stimuli over 3 weeks.

A 3 x 3 repeated measures ANOVA on these means found significant effects of week ( $F[2,24] = 39.6$ ,  $p < .001$ , GG) and the interaction of stimulus and week ( $F[4,48] = 25.9$ ,  $p < .001$ , GG), but the effect of stimulus type, surprisingly, was only a trend ( $F[2,24] = 2.48$ ,  $p = .11$ , GG). It should be noted that these data contain both detected and non-detected subjects in the second and third weeks, and are shown simply to convey the general trend of what was suggested in the ERP figures above. It is clear that the probe response is decreased from weeks 1 to 2, and stays reduced in week 3. The irrelevant response is increased from weeks 1 to 2, then returns in week 3. The target response is depressed in week 2 and is “released” in week 3. (With b-p values, all 3 effects were  $p < .001$ , GG, and the graphed data would look similar to Figure 22).

Post-hoc t-tests comparing probes for just weeks 1 vs. 3 yielded  $t(12) = 4.0$ ,  $p < .003$  (b-p) and  $t(12) = 4.37$ ,  $p < .002$  (p-p). For targets,  $t(12) = 0.66$ ,  $p > 0.5$  (b-p) and  $t(12) = 1.08$ ,  $p > 0.3$ .

In the control group, in 3 x 3 ANOVAs as above, there were significant declines in p-p and b-p P300 across weeks, for Targets and Probes, however as already noted, R remained greater than W,  $p < .001$  in the third week in which BAD still correctly diagnosed guilt in 9 of 10 cases. In an ANOVA on RTs, the only significant effect was a main effect of weeks with a significant linear component suggesting a generalized drop in all RTs over time.

**RELATED, FOLLOW-UP CM STUDY.** With these same subjects, we did another study at a different time. It was the same study with the 6-probe paradigm of Farwell & Donchin (1991) which we ran in STUDY 1 (above). The only difference was the type of subject run. The aim was to compare the difficulty of the 1-probe and 6-probe paradigms in the same set of subjects. It was expected that the 1-probe paradigm would be easier for subjects and therefore result in larger P300 amplitudes with higher detection rates. The results will be briefly and tentatively

presented since there was no control for order. The week 1 results of the previous study will be compared with those of this new study on variables that ought to be related to task demand, e.g., error rate and RT, which should in turn impact target and probe amplitudes. Paired t-tests (2-tailed) are used for all comparisons

For target and irrelevant stimuli, there were no differences in error rates. For probe stimuli, however, there was a significant difference: 3.3% for the 6-probe condition vs. 0.8% for the 1 stimulus paradigm ( $t[13] = 3.21, p < .008$ ). The probe RT in the 6 probe run was 629.14 ms vs. 509.01 ms in the 1-probe run;  $t(13) = 5.3, p < .001$ . The target RT in the 6 probe run was 669.29 ms vs. 557.6 ms in the 1 probe run;  $t(13) = 6.6, p < .001$ . The irrelevant RT was 504.29ms in the 6 probe run, vs. 460ms in the 1 probe run;  $t(13) = 1.73, p = .107$ . These results support the conclusion of greater task demand by the 6-probe paradigm, and predict larger amplitudes to oddballs (probes, targets) in the 1-probe run.

For p-p amplitudes, the probe P300s were greater in the 1-probe run by 3.3 microvolts;  $t(13) = 2.88, p < .02$ . The target P300 in the 1-probe run was 1.6 microvolts greater than it was in the 6-probe run;  $t(13) = 2.32, p < .04$ . The results were similar with b-p amplitudes for probes, but not for targets where the effect did not reach significance (although the 1-stimulus value was .66 microvolts greater than in the 6-stimulus run). The 1 stimulus paradigm detected 12/13 of the subjects (BAD), whereas 10/13 were detected in the 6-probe paradigm. The results clearly support the hypothesis that the 1-probe paradigm has the lower workload.

## GENERAL DISCUSSION

We have shown that the 6-probe, P300-based CIT paradigm can be defeated, and that RT analysis can not provide certain identification of explicit CM users, whose CM responses can be made covert and undetectable: mental or subtly physical. In view of the results of the third week of the 1-probe study, we speculate that defeat of the 6-probe paradigm--like the 1-probe paradigm--might also not even require explicit CMs after having had a practice session with explicit CMs. This is an as yet untested empirical question. We also presented evidence that, as would be intuitively expected, the 6 probe paradigm produces more task demand than the 1-probe paradigm. This extra demand is manifested in elevated error rates and RTs, and diminished probe P300 amplitudes, and thus, lower detection rates. This finding must be offered cautiously as there was no control for order in comparing one set of subjects who performed the 1 and 6-probe paradigms on separate occasions, always in the same order. Yet if the finding holds up, it accounts for the overlapping irrelevant RT distributions in the 6-probe paradigm in CM-users, but not in the 1-probe paradigm: RTs are elevated in the 6-probe paradigm to irrelevant (and probe) stimuli thus pushing the RT distribution of the CM user into the range of either the simple guilty subject or the guilty subject using a CM; (see Figs 7 and 8 and the paragraph following Fig 8).

The 6-probe paradigm has other theoretical difficulties not discussed prior to this report: One assumes that Farwell & Donchin, (1991) chose to use 6 probes because the developer of the GKT (Lykken, 1981) recommended a minimum of 6 items on a polygraph-based GKT. The point of multiple items is as follows. If for one item there is a choice of 5 alternatives, then the probability of a chance hit on that item is  $1/5 = 0.2$ . The use of more orthogonal items reduces the multiplied fractional chance hit probabilities to 0.000064 with 6 items if all items elicit guilty responses. With a 6-item test, even hitting on just 3 items yields  $p = .008$  chance hit probability. The point is that in the format of a standard polygraph GKT, one has responses to *each*,

*individual probe*. This is *not* the case with the paradigms of Farwell & Donchin (1991) or Farwell & Smith (2001) which average all probes together. Let us suppose that a subject produces a consistent P300 to just one and only one of the 4 probes--for whatever reason, including recognizing this one guilty knowledge item. The resulting average ERP to all 4 probes should contain a small P300, as it is an average of 3 actual irrelevants with one probe. The target will reliably produce a large P300. The BC-AD method, as Farwell & Donchin (1991) use it, looks at cross correlations, which will, in calculating correlation coefficients based on standard scores, scale the amplitude differences between probe and target away and likely declare guilt, not knowing which or how many probe items were really recognized. The BAD method might also find the probe greater than the irrelevant and also produce a false positive. The 6-probe analysis method could be modified to consider one probe at a time, but then the numbers of target and irrelevant trials would have to be cut by 75% (to make them comparable with the reduced number of probes), which is likely to result in too few trials for reliable analysis.

We would suggest that the use of repeated blocks of the 1-probe paradigm, with a new probe on each block (and perhaps new sets of targets and irrelevants) is more likely to avoid the problems just described in the 6-probe paradigm. Moreover, in the 1-probe paradigm (unlike in the 6 probe paradigm), at least in our sample, there was no actual overlap of RTs to irrelevant stimuli between the naïve guilty and explicit countermeasure runs. One would have to run many more subjects to confirm the lack of overlap, but it is certainly conceivable that even if there is slight overlap, a cut-off RT could be determined to identify *explicit* CM users. Moreover, in the 6 probe paradigm, the target P3 is substantially larger than the probe and irrelevant responses during explicit countermeasure use, making the pattern of the three ERPs closer to that of the innocent pattern—the classical defeat profile. This was not true in the 1-probe paradigm where all three P300s were similarly small. This alone could raise the suspicion of the investigator, but could not be taken as undisputed evidence of non-cooperation, which can be provided only by behavioral data such as RTs and error rates.

The remaining serious challenge for P300 amplitude recognition indices in deception detection is suggested by the results of the week 3 run with the 1-probe paradigm. There it was found that 5 of 12 subjects still tested (BAD) as innocent without using an explicit CM ;(4 more were undetected with BC-AD). RTs for all subjects in week 3 returned to week 1 levels, confirming the non-use of CMs and also making it impossible to use RTs to help identify former CM trainees. Moreover, the test-beaters in this third week produced classical defeats of the CIT, by presenting ERP averages indistinguishable from those of innocent subjects. Although we do not yet understand the mechanism of this effect, it strongly suggests that the P300 amplitude method may be utterly defeated by a good proportion of those that receive prior CM training. Of course, further research could reveal the mechanism of this effect and allow for more targeted and explicit CM-training. Even if this mechanism were simply due to some kind of habituation effect--which it is not since our control group failed to show the same changes over similar times-- it would still pose a problem for field use as determined test-beaters could practice often with an explicit CM. How, in practice, would this be done?

To practice a CM with a P300-based CIT, a potential or actual wrongdoer would have to have an idea of what probes would be used during the actual test. This might seem difficult in a forensic situation, but in fact, who other than the criminal would recall more real details of the crime scene? He could easily come up with more probes than authorities who can only guess at what should be remembered; the criminal *knows* what he remembers. Now in practice, such a criminal would likely have to consult with an informed professional who has the equipment and

expertise to train the criminal. It is not likely that such a professional (e.g., the present author) would become involved in such marginal activity. So assuming that the scientific community is free of Aldrich Ames types, the forensic situation may be safe. The counter-terrorism scenario is a much different matter. As already noted, Farwell & Smith (2001) and Farwell's web site have strongly promoted the use of the P300 CIT as a counter terrorist tool. Various security professionals were shown to possess concealed information using the P300 paradigm. The generalization is that a member of a terrorist organization has guilty knowledge details about his organization: acronyms, names of lower level leaders, training camp layouts, etc. Assuming our security agencies also have some of these details--which may be somewhat of an assumptive leap—a CIT could be composed for would-be or actual terrorists. In this situation, it is clear that the terrorists certainly can guess well ahead of the test what probes may be used, and so practice the CM preparation. Since these individuals will likely come from a different culture and society whose professional members could be sympathetic with the goals of the foot soldiers, or who could be coerced into cooperating, obtaining professional training might not prove to be difficult.

Finally, we have shown that the method of analysis appears to interact with subject type. Subjects cooperative with experimenters are detectable just as well with BAD as with BC-AD, but truly naïve subjects, which would likely include those encountered in the field, are poorly detected with BC-AD.

## 2. SCALP DISTRIBUTION STUDIES

Because of the problems with use of P300 amplitude as a recognition index in deception detection (as listed in the background section of the introduction) and in view of the success of countermeasures to this method just reviewed, in recent years we have studied the P300 profile or amplitude distribution across the scalp--an "isovoltic" brain map--as an index of deception using mostly GKT analogues. Johnson (1988, 1993) has given elegant theoretical accounts of the significance of varying scalp profiles in which he emphasizes that each particular profile represents a unique pattern of activation of P3 neurogenerating neurons associated with a particular psychological state or task. Thus if two tasks or conditions, within or between subjects, produce differing amplitude distributions (or profiles), one may infer that differentially located groups of P3-neurogenerating neurons are involved by the two conditions. Although there are other explanations for differing profiles (e.g. Donchin, Spencer, & Dien, 1997), all are consistent in assuming that differing profiles in two states means that the brain is working differently in the two states. The profiles are actually scaled scalp distributions, the scaling being necessary to guarantee that the scaled amplitude profiles are orthogonal to simple amplitude differences (McCarthy & Wood, 1985; Johnson, 1988, 1993). The typical familiar statistical method of showing that two group profiles differ at the same  $n$  sites is to do a 2 (groups or conditions)  $\times$   $n$  (sites) ANOVA on the scaled amplitudes and show a significant group by-site interaction.

We have thus recently utilized the scalp profile as another brain wave indicator of the two states of deception versus truth-telling. One of our first studies (Rosenfeld et al. 1999) along this line involved the use of a match-to-sample paradigm with nine probes. In this paradigm, a 3 digit (sample) number is presented on the screen and removed. A few seconds later a probe number is presented and the subject must decide whether or not the probe matches the sample. Of the nine probes presented prior to the next sample, there is only one match. There were two groups of subjects: liars (L) and truth-tellers (T). The T-subjects were told to do their best on the

easy test, and that they would probably score 100% correct. The L-subjects were manipulated to score 50% correct on both matches (MAT) and mismatches (MIS). Thus for T-subjects there were two possible stimulus-responses combinations: 1) Match stimulus and Match response (Mat-Mat), 2) Mismatch stimulus and Mismatch response (Mis-Mis). For the L-subjects there were two additional possible combinations, Mat-Mis and Mis-Mat on dishonest trials. One major finding was that a comparison of T and L subjects' scaled P3 amplitudes as a function of site (Fz, Pz, Cz) and stimulus type (regardless of response) yielded a significant group-by-site interaction, meaning that the P3 profiles of the truth group differed from those of the liar group. Both the Mat and especially the Mis profiles of the liar group showed a quadratic component, whereas the T-profiles appear more linear (or, as we jest in the lab, "liars are crooked.")

The other major finding of this study was that within just the liars, regardless of the stimulus type, the profiles superimposed, suggesting a deception-specific profile. In the truth-tellers, the two corresponding profiles clearly differed, as they should have, since the brain probably does process matches differently than mismatches and thus the associated neurogenerator sets recruited by the two kinds of processing should be different. However, engaging in deception appears to swamp out these effects, as just noted.

As the results just described were being collected, we were prompted to re-analyze profile data from some older published studies in which we had the profile data set but analyzed it only for simple amplitude and latency effects at one site only. One study utilized an autobiographical oddball paradigm with subjects' birthdates as oddballs. The other study utilized a match-to-sample paradigm with one test probe per sample. In each study, there were two conditions, a truth-telling condition and a lie condition in which subjects were instructed to lie on about 50% of the trials. The key results (from Rosenfeld et al., 1998) were that if one looked only at the truth conditions in both studies there was a clear interaction (statistically confirmed) between the autobiographical and match-to-sample studies. This was quite expectable since the two paradigms have obvious differences, requiring differential cognitive processes which should activate differentially located sets of neurogenerator neurons, a situation resulting (see above) in differing scalp profiles. In contrast, the profiles from the two paradigms in the lie conditions yielded no significant interaction and were indeed virtually superimposed. Again, there seemed to be a deception-specific profile which seemed to swamp out other influences which could express their effects in truth-tellers.

In another study (Miller, 1999a; Miller et al., 2002) using an autobiographical paradigm, two groups of subjects, truth-tellers (T) and liars (L), were run in two blocks each. In the first, the autobiographical oddball stimulus was the subject's phone number, and all subjects were instructed to respond truthfully. In the second block where the oddball was the birthdate, the T-subjects responded truthfully, whereas the L-group lied on 50% of the trials. We did this so as to be able to compare honest and deceptive ERP responses. The results were that only liars had a P3 profile which differed from the others (all truth-telling blocks), an effect which was statistically confirmed. When the lie and truth response trials were separately plotted for the L-group, however (from Miller 1999a; Miller et al., 2002), no interaction appeared nor was found statistically. Again, it appeared that within the liar group, the deceptive mind-set from lie trials carried over in truth-telling trials, swamping out the effect of the specific behavioral response of either truthful or deceptive.

Most recently, we have been able in two studies to obtain response-specific profiles during 50% lie blocks (Miller, 1999b, Rosenfeld, Rao, Soskins, & Miller, 2002). In the first of these studies (Miller, 1999b), a novel feature was the addition of four different recording sites on the

scalp to add to the three midline sites (Fz, Cz, Pz) we used previously in all studies. In the other study (Rosenfeld et al., 2002), we utilized verbal responding, rather than button-pressing. We believe *it was important for us to have shown that it was ultimately possible to observe different profiles for honest and dishonest trials within one trial block for both theoretical and practical reasons*: theoretically, this demonstration even more strongly supports the notion that the profile measure may be a direct index of deception. (This is discussed fully in Rosenfeld et al., 2002.) The good theoretical point turns out to have an important practical implication: if there is a direct index of deception, one need not infer deception from the inconsistency of P300 amplitude and behavioral response as one does in the paradigm in which P300 amplitude at one site is used as indirect recognition index and thus deception indicator (e.g., Farwell & Donchin, 1991; Rosenfeld et al., 1988). In the latter, older approach--as with all GKT approaches--unless the test-maker goes to the extreme lengths suggested by Lykken (1981, pp. 257-296) in development of the GKT, which few testmakers are willing or able to do (accounting for the unpopularity of GKTs in the deception detection community), then the absence of a P300 in a guilty person to a relevant item may falsely indicate innocence when the simple truth is that the stimulus item was never noticed or forgotten by the perpetrator. Even more practically, in the type of screening analog which we report on here, it becomes possible to compare responses to various items within a trial block, some of which are answered truthfully, some falsely.

In view of the forgoing, we here attempted to develop a deception detection test based on the P300 profile. The work described below goes well beyond what we have previously reported: 1) We utilized 30 scalp recording sites in the hope of sampling the brain with finer resolution than in our previous work which utilized 3-7 sites. 2) We developed *completely novel* statistical tests designed to be utilized *within subjects* to determine innocence or guilt. All results previously presented regarding profile were based on standard group analysis methods. Moreover, the individual bootstrap tests utilized in the earlier GKT-type paradigms are designed to answer the simple question: is the P300 at Pz in case A larger than that in case B? Our newer approach, in contrast, asks: is the profile shape across 30 sites different in deceptive vs honest conditions? This is a much more complex undertaking and required altogether novel (and commercially unavailable) software. 3) In fact, although all 30 sites were utilized for initial recording, we appreciated (based on Donchin et al., 1997) that there would be redundancy in clusters of correlated sites, and so prior to utilizing the individual diagnostic software, a Principal Component Analysis (PCA) was performed on the set of ERP averages from all sites in all subjects in all conditions. The point of this analysis was to determine a group of *spatial components* or “virtual sites,” each composed of highly correlated real sites which would allow a more efficient application of the diagnostic tests. This is because the PCA reduces the number of points in all analyzed profiles, and incorporates presumably, only the meaningful areas (and their component sites) where deception-relevant activity is present. This is a practical matter, since it should yield more reliable, valid, and accurate detection. Theoretically, identifying these deception-relevant areas gives us an idea of where on the scalp deception-related brain activity is manifest. One cannot generalize from scalp to brain location with great accuracy, but identification of deception-relevant scalp areas gives us a preliminary, if crude idea of what the deception-relevant brain areas are.

## Methods

Subjects: The 23 subjects reported on here (15 males) were drawn from the introductory psychology course pool at Northwestern University. All had normal or corrected vision (but no

contact lenses were permitted). All participated in the research as a fulfillment of a course requirement. Initially, 48 subjects were recruited. Table 4 tabulates the reasons for attrition to the final 23.

**Table 4:** Use of 48 SubjectsGood Files:

Innocent subjects	7
Guilty of one act	7
<u>Guilty of two acts</u>	<u>9</u>
Total	23

Bad Files:

Dismissed due to outside construction noise	5
Data lost due to computer crashes	6
Data dropped due to excessive blinking	2
Data dropped due to sleeping in run	2
Data dropped due to high electrode impedance	2
Data dropped due to excessive head size	1
Data dropped due to excessive coughing	1
Data dropped due to guilt on 5 items	1
No-shows or would not complete experiment	5

(It is noted that over 100 subjects were run previous to the 48 tabulated here. Many were put through a pilot study using an autobiographical oddball paradigm and using 15 sites. These subjects were run to allow us to validate our analysis software. Some of the results are in Appendix 2.)

### Procedures

Subjects entered the laboratory and were informed about the nature of the experiment, and then signed a consent form. An electrocap was now applied with 30 electrodes. (Two other electrodes were used to monitor EOG). Our impedance criterion was 5 kohms. The subjects were then shown a list of eight antisocial or illegal acts to study as an audiotape of those acts was played. The acts on the list were two-word phrases, such as 'smoked pot,' and the full meaning of the phrase was explained on the tape, e.g. "The phrase "smoked pot" means that you have smoked marijuana at least once a week for a three or more month period at some time in the past five years." After hearing the meanings of the phrases, the subjects read aloud sentences both admitting and denying the acts (see Appendix 3). The aim of this was non-selective activation of their memories of their actual acts, as in Johnson & Rosenfeld (1992). The subjects were led to believe we were recording brain activity during the activation phase (we were not), so that just between it and the real recording period we could make the following mock accusation: "We think you did A but possibly also B, C, or D." Of acts A-D--selected to be of moderate probability in our population based on much previous study--we would compare ERPs associated with control (accused innocent) acts with those profiles associated with relevant (accused guilty) acts. The subject then was led into the recording room where the stimuli were presented on a display screen 1m before them. Stimulus items (the 2-word phrases) were presented one at a time every 5 sec. After the recording session, the subject was led to a private room and allowed to shut the door, leaving him/her alone. Next to each phrase on a fresh two-phrase list, were yes and no blank lines. The subjects were told to check one line (yes or no) according to their best memory. They had been led to believe that they were alone in the room and that they could keep the list on them to dispose of as they pleased later. In fact, a closed circuit television system with a concealed camera recorded the list during line checking. (Subjects were later debriefed about this and all procedures as approved by the Northwestern IRB.) The system did not record on tape or anywhere else what the subject checked. It simply carried the signal via cable to a closed circuit TV monitor observed in a neighboring room by an experimenter. Only the list was shown. No identifying information was stored. This was our method of obtaining *ground truth*.

The trial structure was as follows: The trial began with pre-stimulus ERP recording for 104 ms. At 104 ms, at random, one of nine two-word phrases was presented, centered above a fixation point. Eight of these were the phrases seen earlier by the subject. A novel phrase was "Lie Test." We had explained to the subject prior to running that this phrase would be presented and that it meant "you are taking a lie-detector test, which you really are, so you must press this yes button when you see this phrase. Since you want to appear of good character, you will press the no-button to the other phrases even though you may be lying to one or more of them when you do. We want to see if our brain-wave lie-detector can catch you." The subjects rested their right hands on a box containing yes and no-buttons, and responded with index finger presses as appropriate. The stimulus remained on for 1296 ms and the recording continued until 2048 ms. Immediately following the end of the recording epoch, the message "Respond" appeared and lasted 1.5 s. The response had to occur in that window, or the trial was rejected. After it

expired, another 1.5 period of no events occurred prior to start of the next trial. The total intertrial (interstimulus) interval was thus 5 s.

### EEG Methods, Data Processing

Thirty electrodes in an electrocap were applied as shown in Figure 23, based on Pivik, Broughton, Copolla, Davidson, Fox, & Nuwer, (1993). (This figure shows all sites in Pivik et al., 1993. We used only a subset of 30 including those with colors filled in, which indicate component loadings as described below.) The electrocap leads were led to a 32-channel amplifier system from Contact Precision Instruments. The nose was the reference and the forehead was grounded. The amplifiers passed signals from .3 to 30 Hz and amplified them by a factor of 50,000. The outputs were led to a 12-bit Keithley-Metrobyte A/D converter sampling at 125 Hz, which was connected to a Pentium II computer running at 500 MHz. The software was designed to record EEG while presenting stimuli and controlling all aspects of the experiment. After a run of 256 trials, single sweeps were averaged, filtered and compressed for storage on a Castlewood ORB drive. Averages were stored for display as well as for use in a principal component analysis (PCA). The averages to each of the nine stimuli were separately stored, and the filtered single sweeps (3 db point = 4.23 Hz) for all trials, appropriately coded for stimulus and response, were stored.

As in the countermeasures studies described above, EOG was differentially recorded from two electrodes above and below the right eye. They were not one above the other, but diagonally placed so as to pick up both vertical and horizontal eye movements as confirmed in pilot studies. The on-line artifact rejection criterion was 80  $\mu$ V.

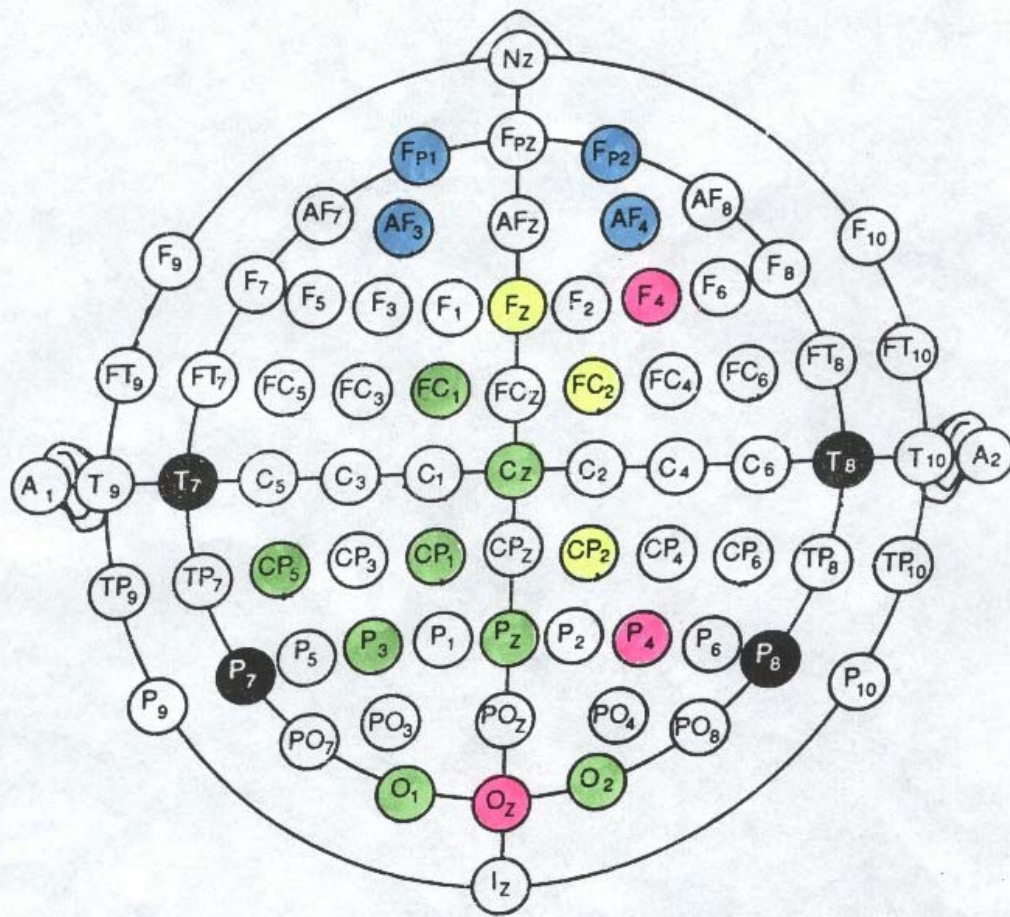
Both stored single sweeps and averages were exported as text files into SCAN 4.11 (Neuroscan Corp). Ocular and other artifacts were removed with SCAN's ocular artifact removal and linear detrending transforms. Then, we visually inspected single sweeps on all files and hand-removed suspicious sweeps, (about 15 per subject). The artifacted averages were then exported to SYSTAT 8.0 (SPSS Co.) where a PCA was performed on 28 of the 30 electrodes. It was decided to drop sites T 7 and F 8 as these lateral sites were found to be quite noisy in several subjects. (This was probably due to flaw either in our electrocap application technique or in the electrocap.) Thus the data matrix input to the PCA consisted of 28 electrode sites by 14,490 observations (70 timepoints X 9 stimulus types X 23 subjects).

The purpose of the PCA in space was to tell us how to use the data from the 28 sites actually recorded. That is, the PCA tells us which sites contain mutually redundant information. All these can be put into a cluster or virtual site containing information from its component sites which the PCA also shows to be good representations of a common factor or component. The PCA provides that information by outputting a set of "loadings" for all recorded sites. Only some of the actual sites will load highly on (i.e., represent strongly) a given factor.

Once we completed the spatial PCA, we computed virtual ERP single sweeps which are weighted averages of site values within spatial clusters. We then performed statistical analyses (described below) on *each subject* using bootstrapped cross-correlations and ANOVAs. These involve many interactions and are very time consuming (even on a 1.7 GHz Pentium 4 which we used for iterative analysis). Thus, once a PCA is run, its results shape the analyses on all subjects. It took 2 weeks to do the PCA and intra-subject analyses on 23 subjects. Actually, 2 PCAs were run, one with a varimax rotation and another with an oblimin rotation, (as suggested by reviewers of first draft of this report).

There were basic choices to be made in terms of how to run the PCA: How many factors (virtual sites) should be used? What should be the criterion for dropping real sites from virtual clusters? Should sites within a cluster be weighted (multiplied) by respective loadings, or by 1.0? The component loadings are shown in Figure 23 (varimax orthogonal rotation) and Figure 24 (oblimin non-orthogonal rotation). As the outputs of both rotations looked similar (and substantially different from the PCA outputs submitted prior to the first peer review, we chose to use the orthogonal rotation, as initially proposed, and as is typically used by ERP workers.. Table 5 lists the 30 sites used in the study and the virtual sites on which they loaded significantly, if any. Specifically, we chose for this report to go with 4 factors or virtual sites. This decision was based on PCAs, and most especially by the Scree Plot Test (Catell, 1966) results. This plot of eigenvalue as a f (factor number) for the presently submitted dataset is shown in Figure 25 for the oblimin rotation; the varimax was similar. This plots essentially the amount of variance in the data accounted for by the various factors from 1 to 28 (the number of sites/variables). It is seen that there is one key factor with eigenvalue > 200,000. There appeared to our eyes to be three more real factors, after which all points seemed to be on the same trend line. Our PCA on four factors accounted for about 74% of total variance, clearly a sizeable amount according to our factor analysis consultants. These four factors accounted for about 31.8% (factor 1), 17.5% (factor 2), 15.4% (factor 3), and 8.9% (factor 4) of variance, respectively. Other PCAs loading sets, using different numbers of factors were done. These were covariance-based PCAs. Of course, the more factors one uses, the greater the total amount of variance accounted for, but there is a point of diminishing returns: where there are no further inflection points in the Scree Plot. We utilized a correlation matrix PCA to obtain standardized rotated loadings for each of four factors. We then decided to retain all sites with loadings of .5 or more (in a 0 to 1.0 range). This is a compromise choice, as with the number of factors. Thus for each of the four components (factors) we identified, all sites showing loadings of .5 or more were simply averaged to give us a value for the virtual site (cluster or factor). We might have chosen to use a weighted average, with each site value multiplied by its respective loading, but we reasoned that such loadings differences (given our .5 cutoff) could be due to chance and our results would then not generalize to other datasets. Table 5 shows the loadings of sites on factors.

Figure 23 suggests four areas on the scalp where clusters exist. Virtual site or Cluster #1 (green sites) has a large central-parietal contribution. Cluster #2 (blue sites) appears frontal. Cluster #3 appears right frontal-central, and Cluster #4 involves frontal F4 to P4 to OZ. For each real site within a cluster, the filtered single sweeps for each real site were converted to a set of filtered single sweeps for virtual sites by averaging the real site values into one virtual site value for each time point in the single sweeps. Now we proceeded to compare scalp distributions, scaled and unscaled, across comparisons of interest, e.g., control vs relevant distributions within innocent and guilty individuals. To compare scaled scalp distributions, we used a modification of the method of Srebro (1996). This method basically asks if there is a cross-correlation across the two plots (e.g., control and relevant) of scalp amplitude as a function of *virtual* site for the two item types. The use of the Pearson Correlation Function accomplishes the scaling. In our novel approach, we randomly selected with replacement a sample of trials from both control and relevant item sweep sets (separately). We then calculated the average P300 value separately at each virtual site for the selected control and relevant sweeps. Now the actual cross correlation,  $r$ , was computed. After 100 iterations of this process, an average of these  $r$  values was computed, called  $r_a$ . We then took the two original single sweep sets and shuffled them together, randomly



Varimax rotation

31.8%

17.5%

15.4%

8.9%

Fig. 23

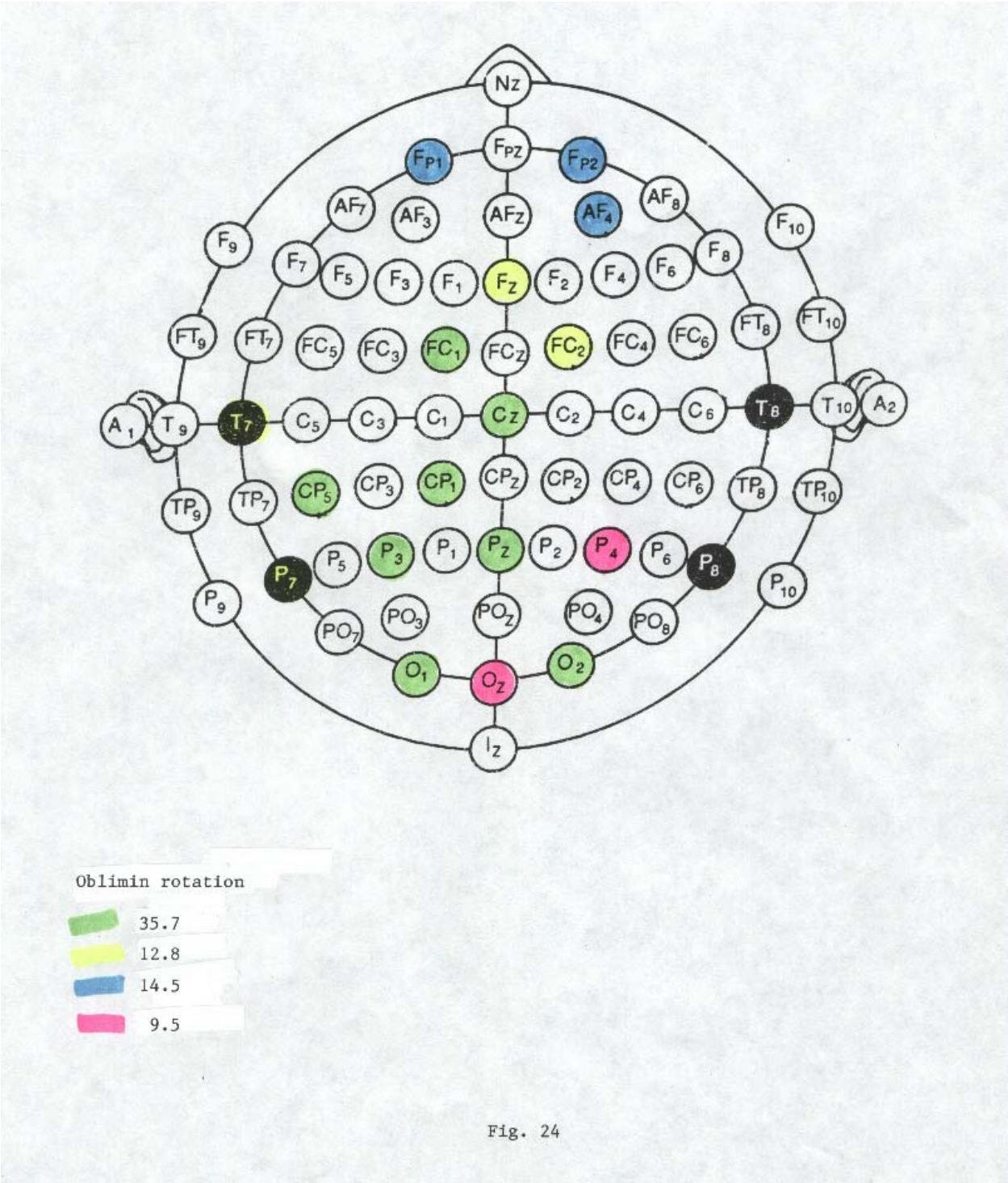
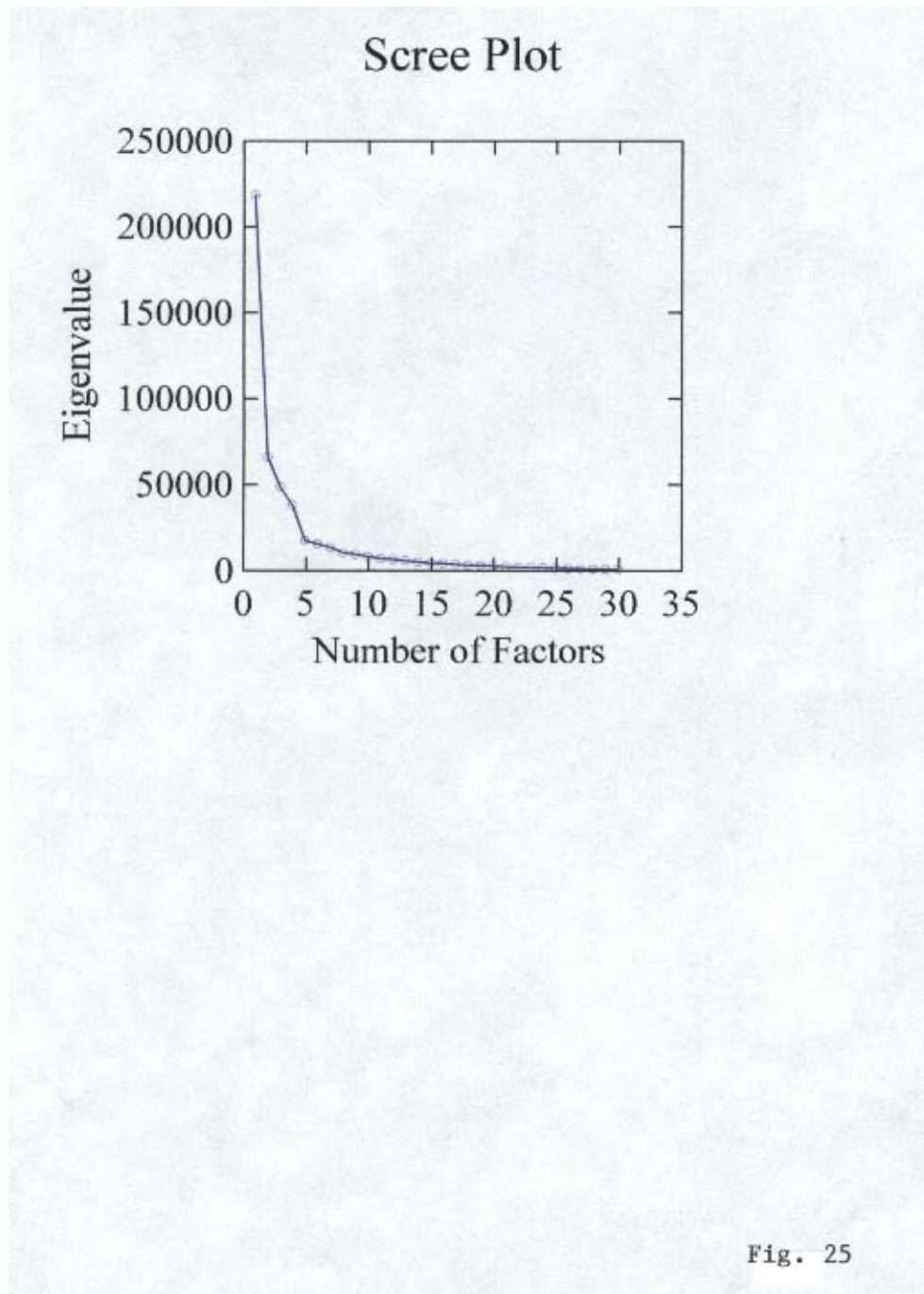


Fig. 24



**Table 5:** a list of the 30 sites used in experiment 3 and the virtual sites on which they load significantly (if any)

site	virtual site
FP1	3
FP2	3
AF3	3
AF4	3
F7	
F3	
Fz	2
F4	4
F8	
FC5	
FC1	1
FC2	2
FC6	
T7	
C3	
Cz	1
C4	
T8	
CP5	1
CP1	1
CP2	2
CP6	
P7	
P3	1
Pz	1
P4	4
P8	
O1	1
Oz	4
O2	1

Virtual site 1 accounts for 31.8% of the variance in the data, virtual site 2 for 17.5%, virtual site 3 15.4%, and virtual site 4 8.9%.

combining control and relevant sweeps. We now arbitrarily “cut the deck,” i.e., we divided the shuffled set, and treated the first fraction of single sweeps as one (pseudo) condition and the second fraction as another. The fraction numbers corresponded to the actual numbers of single sweeps per condition. As previously (with unshuffled data), sweep sets were repeatedly drawn with replacement from each condition, averaged, and the iterated ( $n=100$ ) cross-correlations ( $r_b$ ) computed. In order to conclude that the subject was guilty of the tested item, the  $r_A$  value needed to be in the *lower* 10% tail of the distribution of  $r_b$  values.

For the unscaled data we used an ANOVA approach. (Although we had noted in the August 2000 progress report that the ANOVA approach wasn’t working, we most recently found a correction and the program now does what it is supposed to do.) The plot of P300 amplitudes as a function of virtual site, separately for control (B) and relevant (A) data, lends itself to a 2 (condition) by 4 (virtual site) ANOVA. The approach taken is bootstrapping as with the correlation method, except instead of calculating  $r$ , the cross-correlation, we perform an ANOVA on each set of iterated site-condition values (8 in all =  $4 \times 2$ ). The F-term for interaction states whether the curves from the two conditions are parallel or not. If not, the conclusion is the two profiles differ. The specific criterion here was that the  $F_A$ -average on real iterated values be in the top 10% tail of the distribution of shuffled  $F_B$  values.

More detailed summaries of these methods are given below.

## RESULTS

The stimulus categories used here were as follows: 1) Relevant items were those accused items ( $n$  = either 1 or 2) of which guilty subjects were guilty. 2) Control items were the *falsely accused* items of which subjects were innocent. 3) Target items were the “Lie Test” items (see above). 4) Irrelevant items were the remaining innocent items *not* used in false accusation. They have a low probability in our student population (e.g., “ROBBED BANK”).

## INTRA-INDIVIDUAL DIAGNOSES WITH SCALP DISTRIBUTIONS:

This major goal of the originally proposed project was to explore the use of scaled and unscaled scalp distributions as diagnostics of deception. We first insert here a brief summary of the analytic methods used and their acronyms:

### Analysis of Scalp Profile (ASP)

This analysis method to be used in these studies is called the analysis of scalp profile (ASP) method. ASP uses a stimulus type by site amplitude interaction analysis as the basis of a decision about guilt or innocence for each subject. One can determine the average amplitude at each virtual site for each stimulus type, and plot pairs of curves. The basic question being answered by ASP is, “Are these curves parallel?” It is expected that for innocent subjects, the control and designated “guilty” or “relevant” stimuli will have parallel curves, whereas for guilty subjects these two stimuli will not yield parallel curves for both scaled and unscaled data. Rather than simply employing an analysis of variance (ANOVA) on the actual raw single sweep data, which are noisy, ASP performs an ANOVA on bootstrapped averages. This ANOVA is not by itself performed to determine significance, but the resulting F value is instead compared with the iterated F values resulting from the second part of ASP. In the first part of the ASP procedure, the actual sample with replacement bootstrapping procedure on real data used in ASP to obtain

bootstrapped averages is similar to that used in BAD (see above). From each set of  $n$  single sweeps from a particular stimulus,  $n$  single sweeps will be randomly drawn with replacement. The same is done at all virtual sites for both stimulus types to be compared. This process is repeated 100 times yielding the number of virtual sites times 100 sets of individual, bootstrapped average P300 values. Now one has a sample size of 100 for each site/item combination and the within individual ANOVA can be carried out on the bootstrapped P300 averages. This is done 50 times and the resulting F-values are averaged. The resulting average F (F-real) is compared with the F-values resulting from part two of ASP.

The only difference between the second part of ASP and the first is that in the second part, prior to random (with replacement) selection of single sweeps to be averaged, the data from both stimulus types are shuffled so that during subsequent random selection, single sweeps are drawn from a set that probably contains both stimulus types. After this procedure, the same ANOVA as performed in part one is performed on the shuffled data to yield an F-value (F-shuffled) for two pseudo groups in which real data were shuffled together thus destroying any real difference between groups (we will call this the base rate). This procedure is now itself performed 100 times to yield a distribution of size 100 of base-rate F-values to which the F-real F-value obtained from part one is compared. If the average F-real value obtained in part one is greater than 90% of the F-shuffled values obtained in part two, then a guilty decision is made, whereas an innocent decision requires that the part one F-value be lower than 40% part two F-values.

ASP will be used on both scaled and unscaled data separately. When performed on scaled data, the results from ASP are orthogonal to those analyzed by BAD because effects of amplitude are removed by the scaling process. The scaling procedure used is the vector length method (McCarthy & Wood, 1985), adapted for individuals. Ordinarily, in group studies, one divides, within a stimulus type (or condition) the average voltage at one site for a subject by the square root of the sum across sites of the squares of mean voltage averages across subjects. Within one subject (for ASP), one divides the voltage for each (virtual) site on each single trial by the square root of the sum of that subject's squared average (across trials) voltages across (virtual) sites.

### Correlation Analysis Technique (CAT)

The other intra-individual analysis method for scaled scalp distribution is called the correlation analysis technique (CAT). Its purpose is to test scalp distribution differences as in ASP, but using cross-correlation of the scalp distributions from the two stimuli type conditions, e.g., relevant vs. control. Note that the data are automatically scaled by the correlation process. Part one of this program calculates the actual cross-correlation associated with P300 values of two different stimulus types across sites. This yields a real correlation coefficient ( $R$  or  $R$ -real) between the two scalp profiles. A guilty subject would likely have a low  $R$  when distributions associated with control stimuli are compared with distributions associated with the guilty/relevant stimuli, whereas an innocent subject would have a high  $R$  because there is no difference between the designated "guilty/relevant" item and the control stimuli. In order to test the significance of the value of  $R$ -real in a given subject, part two of CAT uses a similar procedure to that used in part two of the ASP program. All of the single sweeps from the control and relevant stimuli are shuffled together and then this shuffled set is arbitrarily divided into two pseudosets of sweeps, analogous to cutting a deck of cards. These pseudosets are each separately averaged, and their cross-correlations are then computed. This procedure is done 100 times to create a distribution of 100  $R$ -shuffled values to which to compare the  $R$ -real value

obtained from the actual data. For a guilty judgement, the real R value should be in the bottom 10% of the distribution of Rs resulting from shuffled data.

#### EXPECTATIONS:

1. It is expected that in guilty subjects, the scalp distribution for relevant responses will differ from the scalp distribution for control responses. For innocent subjects, there will be no such differences. This is a wholly novel prediction based on the assumption that truth-telling and deception are two different cognitive states which will differentially engage P300 neurogenerators which will produce differing scalp distributions.
2. It is expected that the simple P300 amplitude at the virtual site containing the major parietal contribution will be larger in response to relevant than in response to control items in guilty subjects, not in innocent subjects. This is based on Rosenfeld et al. (1991) and is examined here simply as a manipulation check.

#### RESULTS-1

We expect the relevant distribution to differ from the control distribution in guilty persons, but not in innocent persons because for the latter, there are no guilty acts, and therefore, no real relevant items. They can be randomly selected from among the four falsely accused items and designated as relevant.

The nine stimuli used in this study are shown in tabular form (Table 6) below with their symbols, categories (possible Relevant or Control, R/C, or Irrelevant, IR, or Target, TR). and meanings(under ACT):

**Table 6** (see text).

<u>ACT</u>	<u>ACTUAL STIMULUS</u>	<u>CATEGORY</u>	<u>SYMBOL</u>
Stole friend's money	Friend's Money	R/C	F
Robbed a bank	Robbed Bank	IR	S
Used false I.D.	Fake I.D.	R/C	I
Broken store window	Store Window	R/C	C
Took school records	School Records	IR	T
Plagiarized a paper	Plagiarized paper	R/C	A
Smoke pot weekly	Pot Weekly	IR	P
Stolen a bicycle	Stole Bike	IR	B
Taking Lie Test	Lie Test	TR	L

Recall that we did not know in advance of which guilty act(s) if any the subject would be guilty, and as these differed across subjects, any of the falsely accused items (see methods) could be relevant or guilty in any particular case; hence the designation R/C for the four falsely accused acts. Below (Table 7) is a tabulated distribution of outcomes using symbols from above (any symbols not appearing here indicate acts of which all subjects were innocent):

**Table 7** (see text)

<u>Guilty Item(s)</u>	<u>Number of subjects</u>
P and I	4
F and I	3
A and I	1
I	4
F	1
B	1
A	1
None (innocent)	7

Preliminary analyses revealed that there was no significant difference between amplitudes or distributions (scaled and unscaled) for accused versus non-accused acts of which subjects were innocent. Thus one could simply compare the scalp distribution from a subject for all his guilty/relevant acts (if  $n > 1$  or just the single guilty average if  $n = 1$ ) with the average of all remaining, non-target acts. The results of such comparisons (*comparison set 1*) using only b-p amplitudes as originally proposed for ASP and CAT, along with BAD results (p-p) as a manipulation check(see EXPECTATION 2 above), are tabulated as follows:

**Table 8a:** number and percentage of **guilty** decisions.

<u>analysis method</u>	<u>guilty subjects</u>	<u>innocent subjects</u>
BAD (p-p VS1)	13/16 (81.3%)	1/7 (14.3%)
BAD (p-p, Pz)	14/16 (87.5%)	1/7 (14.3%)
ASP (scaled)	4/16 (25%)	1/7 (14.3%)
ASP (unscaled)	7/16 (43.8%)	2/7 (28.6%)
CAT	6/16 (37.5%)	2/7 (28.5%)

**Table 8b:** number and percentage of **innocent** decisions in experiment.

<u>analysis method</u>	<u>guilty subjects</u>	<u>innocent subjects</u>
ASP (scaled)	2/16 (12.5%)	1/7 (14.3%)
ASP (unscaled)	0/16 (0%)	1/7 (14.3%)
CAT	2/16 (12.5%)	0/7 (0%)

The following applies to both tables 8a and 8b: Using McNemar's test of differences between correlated proportions, for the guilty subjects, BAD (peak-peak at site Pz) significantly outperformed unscaled ASP ( $Z = 2.33$ ,  $p < .02$ ). There were no significant differences in correct detections between scaled ASP, unscaled ASP, or CAT ( $p > .05$  for all). Over all analysis methods, there were no significant differences in false positive rates, number of correct innocent decisions, and number of incorrect innocent decisions ( $p > .05$  for all).

It is noted that for ASP and CAT, analyses were restricted to virtual sites only. It is also seen in Table 8a that BAD (Bootstrapped amplitude difference, relevant vs control, as in earlier countermeasure experiments, using a .95 confidence level) was performed on Virtual Site 1, VS1, and actual site Pz, and the hit rates were in the 80%-87% range, as we and others usually see in oddball GKT paradigms. This verified that in terms of simple amplitude, the manipulation worked; guilty but not innocent items evoke P300. The result here, however, has further importance in that it shows that this method will work whether subjects are guilty of 1 or 2 items. This was an important and novel finding, not shown before and it will be amplified later with other comparisons of responses to stimuli where the hit rates get into a respectable range using virtual sites, but not with Pz. (We will also argue that preparing a CM for this screening paradigm is not as simple as with a simple GKT.) Using the *present comparisons*, however, the *scalp distribution* test results in Table 8a are disappointing. We looked further: T-tests performed on the F-values obtained from part one of ASP for both scaled and unscaled data verify that part one of ASP was in fact yielding higher F values for guilty as opposed to innocent subjects *only with unscaled data* ( $t = 6.35$ ,  $p < .001$  for unscaled,  $t = 1.34$ ,  $P > .1$  for scaled). This suggests that the success in distinguishing guilty and innocent groups with unscaled data is based on amplitude confounded with distribution, not distribution itself. Moreover, group effects, however interesting, do not change the low detection rates in Table 8a. One problem with these comparisons, however, is that one is typically comparing a relatively low number of sweeps (30-60 for one or two guilty items)) with a large number of sweeps (180-210 for 6 or 7 innocent items). Then there is the potential confound of accusation with guilty/innocents, despite our lack of finding of such an effect. The following approach solves these potential confounds.

## RESULTS-2

It is noted that prior to undertaking the remaining data analysis, one of the data files from a subject guilty of two acts became corrupted, leaving us 15 guilty subjects (from 16 above), 8 (vs. 9 above) guilty of two acts.

We also did relevant vs control distribution comparisons (*comparison set 2*) in which distributions for accused relevant/guilty items( 1 or 2) were compared with only accused innocent items(1 or 2), or in which non-accused relevant/guilty items were compared with non-accused innocent items. In the case of the four subjects (see Table 7) guilty of P (not-accused) and I (accused) we compared the combined P and I data with combined B(not accused) and C(accused) data. In the case of innocent subjects, we arbitrarily designated two falsely accused items as “relevant” and used the two other falsely accused items (properly) as controls. Thus, all comparisons in Set 2 are unconfounded by accusation and non-accusation, nor by numbers of trials involved, as they may be in Table 8 above. Table 9, below, has the results, in terms of proportions of guilty calls using the CAT analysis on automatically scaled data; (BAD analysis with Set 2 data is considered below under “Simple Amplitude Effects.”)

**TABLE 9.** Proportions of guilty decisions by CAT algorithm in innocent and guilty groups. ”b-p” refers to base to peak amplitudes, “p-p” to peak to peak amplitudes, and “b-p or p-p” designates the numbers of guilty calls ( at 90% confidence levels) using either b-p or p-p analysis.

<u>GUILTY GROUP</u>			<u>INNOCENT GROUP</u>		
b-p	p-p	b-p or p-p	b-p	p-p	b-p or p-p
6/15	8/15	(11/15)	0/7	0/7	0/7
(0.40)	(53.3)	(73.3)	(0)	(0)	(0)

In the guilty group, the superior p-p method yields only 53% detection with no false positives. However, if one goes to the slightly more liberal “either-or” criterion, the hit rate rises to a slightly higher (and respectable) 73% detection rate, without any increase from 0% false positives in innocents. A Fisher exact test comparing “either-or” proportions in Guilty vs. Innocent subjects yields  $p < .05$ .

#### GROUP DATA

These are not directly helpful in terms of practical detection of deception, but may help in our understanding of other results. Thus the analyses to be now described are based on comparison set 2 in which relevants and controls are compared in unconfounded tests. The results below in Table 10 are the outputs of 2-way ANOVAs, completely repeated measures, on the factors site ( $n=4$  virtual) and stimulus type (relevant vs. control) in the guilty group.

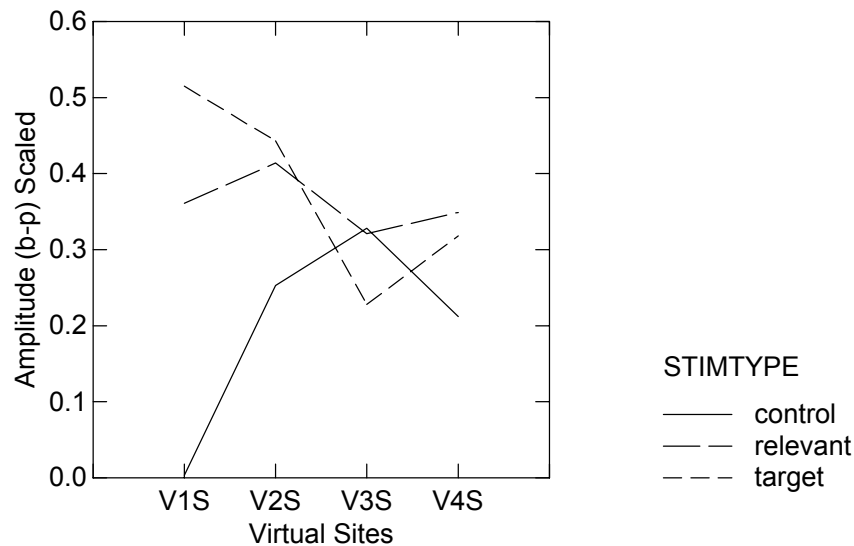
We will present results on scaled data first, and thus look only at the interaction term, as it is the index of distribution differences unconfounded with amplitude. (With  $n=7$  for the innocent group, it was felt that the concomitant lack of power precluded such test results in that group.)

**TABLE 10.** These are F and associated P-values for the type by site interactions in the guilty group.

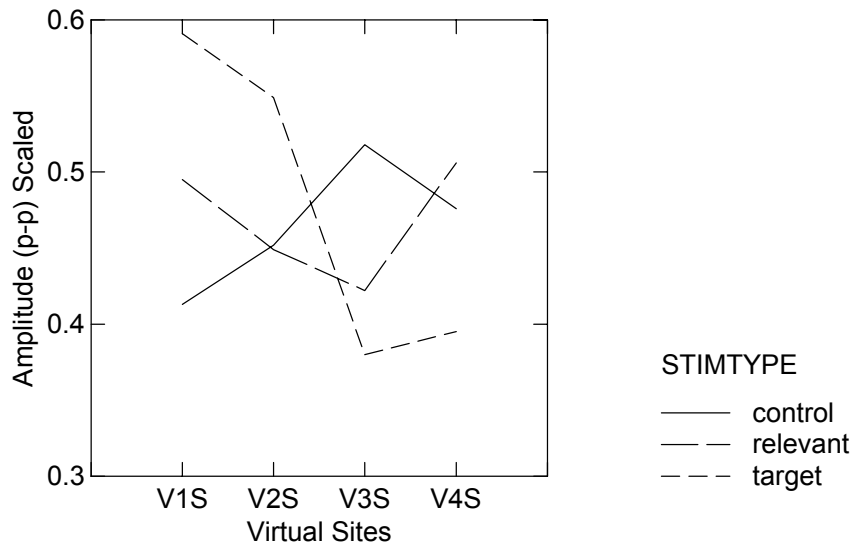
B-P:  $F(3,42) = 2.143$ ,  $P(\text{Greenhouse-Geisser}) = .135$

P-P:  $F(3,42) = 2.11$ ,  $P(\text{Greenhouse-Geisser}) = .137$

These values are only at the trend levels, not full significance levels, although for p-p, the Wilks' Lambda, Pillai Trace, and H-L Trace multivariate tests all yielded  $F = 4.58$ ,  $p < .03$ . (These tests for the b-p values yielded  $p < .09$ .) These scaled data (plus the target values) are plotted in the figures below, one for the scaled b-p values, the other for p-p values



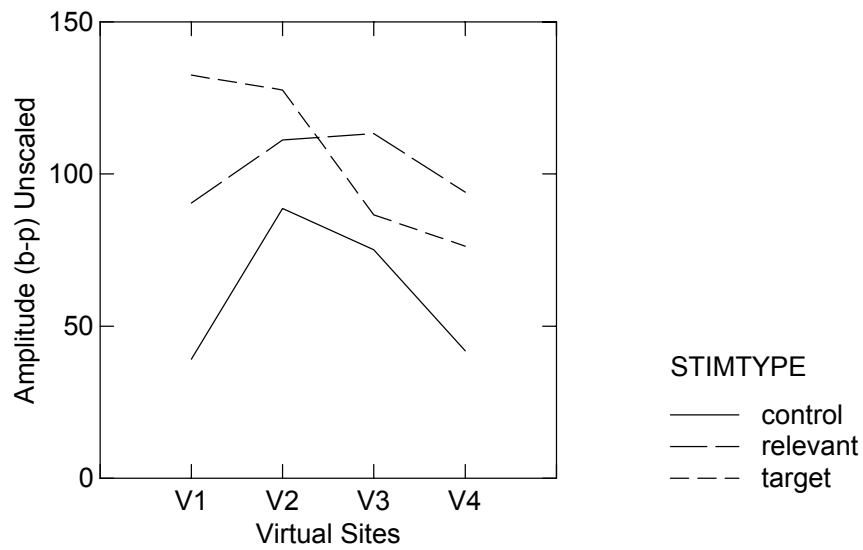
**Figure 26:** Scaled b-p, group-averaged values of 3 stimulus types at 4 virtual sites. These scaled data are shown not to be interpreted, but just to indicate parallel vs. non-parallel curves.



**Figure 27:** Scaled p-p group-averaged values of 3 stimulus types at 4 virtual sites. These scaled data are shown not to be interpreted, but just to indicate parallel vs. non-parallel curves.

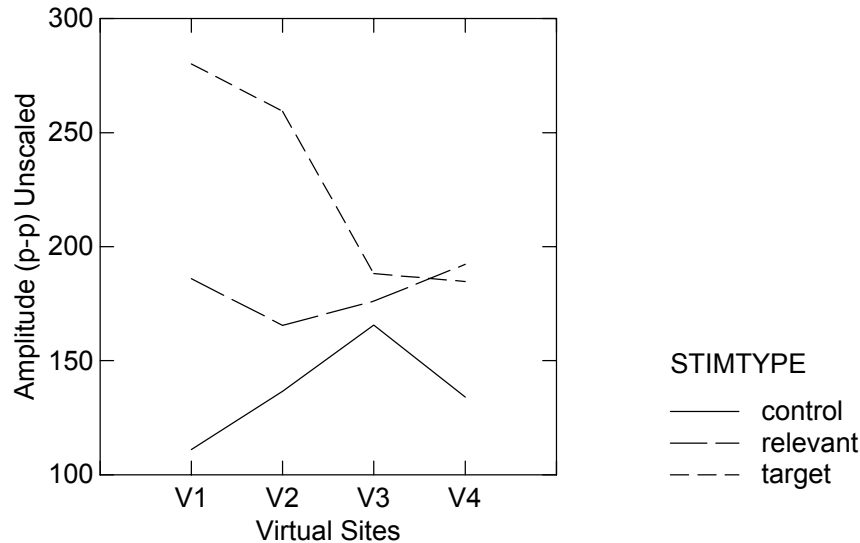
The relevant and control curves in Figure 27 look indeed like mirror images and one would have expected an interaction—which one does get in multivariate tests. However, the marginal nature of the results overall suggests, as we already know, that not all subjects are detected with just b-p or just p-p data. It also suggests a considerable amount of non-systematic or noise effects in the data. It can be added without showing the individual distribution data that *even in those subjects detected with CAT, there was no typical guilty item  $\times$  site interaction, and there were perhaps only 3-4 (of 15) individuals showing distributions resembling those of the group, as above. There is thus no uniform specific lie distribution seen here.*

As a segue into the next section of the results, we now show figures similar to Figs. 26 and 27 on comparison set 2, except that these are for unscaled data. Thus one can here observe type and site effects, although the interaction effect may confound distribution with amplitude (type) effects.



**Figure 28:** These are the group's average coded actual b-p amplitudes ( $10 \mu\text{V} = 409.6$  units) as a function of virtual site and item type.

In Figure 28, above, the main effect of type (Relevant vs. Control amplitude) was  $F(1,14)=3.74$ ,  $p = .074$ . The site effect was not significant, nor was the interaction (both  $p > .3$ )



**Figure 29:** These are the group's averaged coded actual p-p amplitudes (10 uV = 409.6 units) as a f(virtual site, item type).

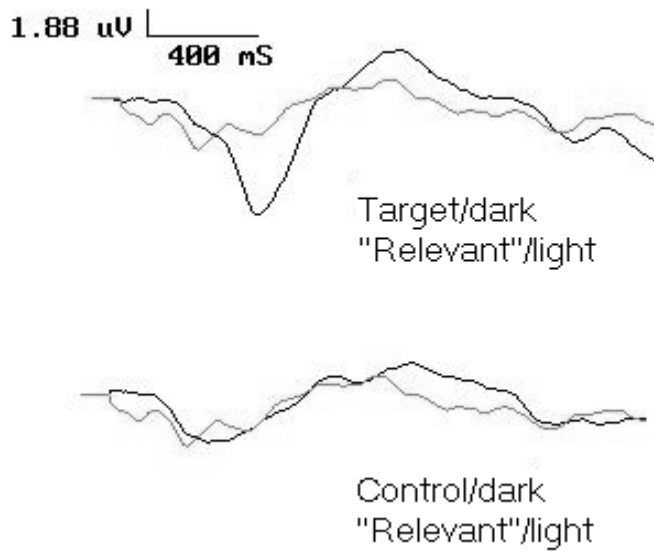
Here (Figure 29), the effect of type was  $F(1,14) = 5.525$ ,  $p < .04$ . The effect appears maximally carried by the first virtual component, the centro-parietal one. There is no effect of site, which appears to interact with type (although the  $F[\text{interaction}] = 2.124$ ,  $p = .123$ . This is marginal, but the three multivariate tests on the interaction were all  $p < .05$ .)

In the main effect of item, we have the classic oddball amplitude effect in which P300s to rare relevants are bigger than those to frequent controls. This is clearly shown in the grand average ERP figures opening the next section.

#### SIMPLE AMPLITUDE EFFECTS:

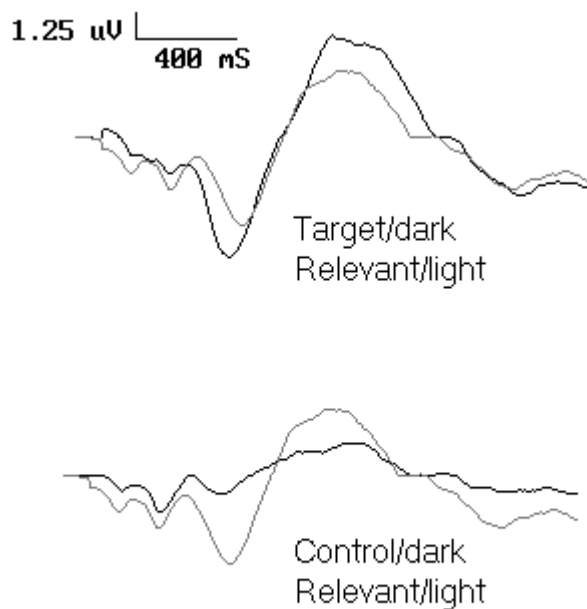
As noted above, the following data do not pertain to the scalp distribution studies except as a simple manipulation check. However, a novel finding emerged, worth reporting.

The following figures show for Virtual Site 1 superimposed grand averages for innocent and guilty groups. For each group, targets and relevants are superimposed, as well as controls and relevants. These averages are based on comparison set 2, where some relevant averages contain one item, and others contain two items, and the comparisons are always symmetrical. First, we have the innocent group. Here there are no real relevant stimuli, so two falsely accused items are designated as "relevant." The other two falsely accused items are combined into a control. It is obvious that control and "relevant" are the same, but target towers over "relevant" (as it would over the similarly small control.)



**Figure 30.** Superimposed grand averages at virtual site 1 in innocent group.

The following figure shows superimposed grand averages as in Figure 30, but for the guilty group. It is evident that the target and the relevant, both being real oddballs, both show similar P300s. It is also evident that the relevant item towers over the control.



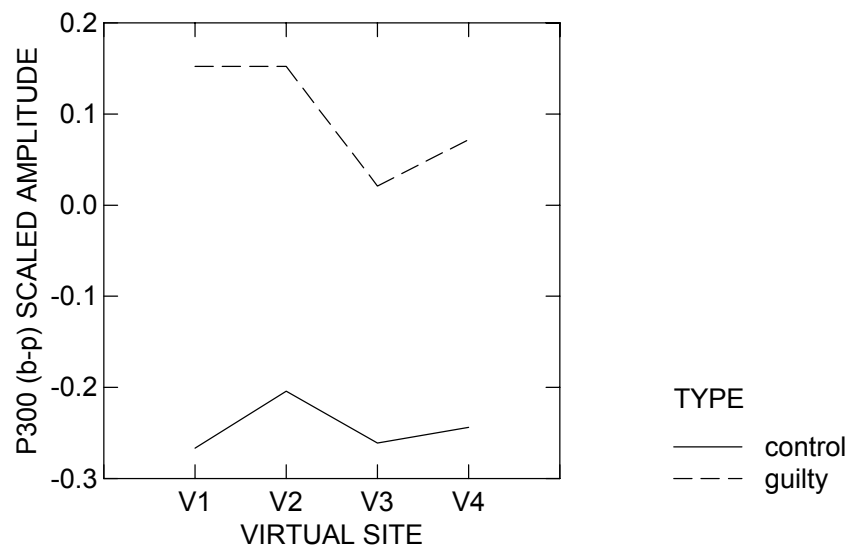
**Figure 31:** Superimposed grand averages at virtual site 1 (the mostly centro-parietal site) in the guilty group. (Note that what we here designate as relevant items are functionally similar to the probes of the countermeasure studies, and the controls here are analogous to the irrelevants of the CM studies. Thus Figure 30 here is comparable to Figure 3 and 4 earlier, and Figure 31 here is comparable to Figs. 1 and 2 above.).

Using the BAD method (90% confidence) on V1 only to test within individuals whether relevant is greater than control, 12 of 15 (80%) guilty subjects were correctly classified and there was 0% (0/7) false positives among the innocents. *Multiple relevants made no difference; i.e., 8 of 8 subjects guilty of 2 items were detected as were 4 of 7 subjects guilty of only one item. This is the first time this has been reported using a more natural screening scenario in which subjects could be guilty of more than one item.* (In the other uses of this protocol, subjects were controlled to be guilty of only one item; Rosenfeld et al., 1991; Johnson & Rosenfeld, 1992.). We were surprised at this unexpected and unpredicted finding and wondered whether or not simple site Pz would do as well as V1, a virtual site containing much but not exclusively parietal contribution, and certainly more parietal signal than just that from Pz (see Fig 23, above). Thus we re-did the comparison set 2 data using just the Pz site (as is typically done in P300-based GKT protocols, such as Rosenfeld et. al., 1991, Farwell & Donchin, 1991, Johnson & Rosenfeld, 1992, Allen, et al., 1992). Pz alone afforded poor detection, correctly naming just 6 of 15 (40%) guilty subjects, only 2 of which were guilty of 2 items. This suggests that even for simple amplitude methods, it is worth the trouble to do the PCA so as to determine virtual sites, and is not consistent with Farwell's statements that additional sites don't help (Farwell & Donchin, 1991; Farwell & Smith, 2000). Perhaps using arbitrarily chosen sites (i.e., without the benefit of the PCA) more sites don't help with only one guilty item, but apparently with two (or more?) such items, the additional contributions to the major virtual principal component do help. Of course this novel effect needs looking into and replication.

As a way of beginning this investigation, we generated a third comparison set. From each of the 8 subjects guilty of 2 items, we arbitrarily selected one. For the other 7 guilty subjects guilty of one item, we used that 1 item. For a control, we chose one falsely accused innocent item. Now we repeated BAD (95% confidence) on V1 and on Pz. For V1, 14/15 were detected (93% hit rate). For Pz, only 9/15(60%) were detected including 5/7 guilty of one item, 4/8 guilty of two items. These results are indeed consistent with the notion that as the number of guilty items within one run increase, the virtual site procedure is increasingly helpful.

#### VALIDITY OF METHODS:

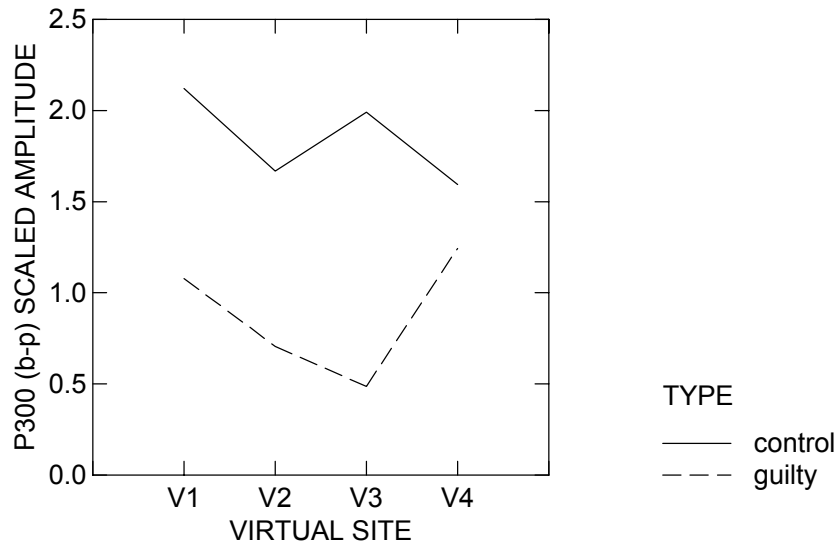
In terms of seeing whether or not the intraindividual analyses results confirmed what the eye saw, we finally present some individual graphs showing a visually obvious interaction, and one with obvious lack of interaction.



**Figure 32:** This is from an individual whose relevant (guilty) item distribution did not appear to differ markedly from control item configuration.

It is noted that in all cases where curves were like Figure, 33 (below, containing an obvious interaction), the CAT test found the statistical evidence of interaction. In cases like Fig 32 (above) where there was clearly no interaction, the CAT program reliably did not detect one. This was true of all subjects in which visually clear outcomes(positive or negative) were obtained, allowing us to comfortably rely on the CAT program for determinations in those other subjects whose outcomes were not visually obvious. (It is noted that these figures are based on

the comparison set 2 data in which the comparisons were controlled in terms of accusation; i.e., accused guilty/relevant items were compared only with accused innocent/control items, etc.)



**Figure 33**

### DISCUSSION: Scalp distribution study

Regarding the scaled scalp distribution, or profile method of detecting deception, it is not clear from this study how useful the method may become. We saw that by using either a positive outcome from CAT on b-p or p-p profiles, we could detect 73% of the guilty subjects, with 0% false positives in innocent subjects tested with the same criterion. These results were based on the least confounded comparisons within subjects regarding accused versus non-accused items. That is respectable, but needs replication and improvement. More importantly, this method requires more systematic investigation with respect to guilt criteria used in the CAT algorithm as well as effects of accusation. Time did not permit us to systematically vary confidence level in this study. That is, for a guilty decision, the CAT method we use requires( as rationalized above) that the real cross correlation coefficient between relevant and control items across sites be in the bottom 10% of the distribution of shuffled, bootstrapped cross-correlations; i.e., we operate at the 90% confidence level. Perhaps this is too stringent. Perhaps the CAT method is so sensitive that if we dropped the confidence level to 80% or less, we would begin approaching 90 to 100% detection of guilty subjects with little or no cost in false positive rate. We did not see the point in pursuing this with the present data set because we had only seven innocent subjects, and because the guilty subject set contained a highly variable set of guilty acts per subject (see Table

7 above), some of which were accused, some not. This was the result of our electing to use a very naturalistic screening scenario of the type the government might find useful.

In future work, in view of the promise seen here with appropriate, unconfounded comparisons (*set 2*), it will be worthwhile to better control the accusation factor. In this regard, it must be mentioned that the demographics at this university must have changed since we last used this protocol 20 or more years ago. At that time, the item “used false ID” was one of which 50% of the student subjects could be counted upon to have committed. In contrast, none of the students in the 1980s reported stealing friends’ monies. Sadly, this has changed. Some careful demographic piloting next time might allow us choose a list of items resulting in more predictable guilt profiles.

We conclude that the profile method is worth pursuing, because it does not seem as if there can be a simple countermeasure made readily available. It was self evident to our laboratory from the earliest days of P300 based detection of deception with P300 as a recognition index in GKT paradigms, that it would be possible to covertly make irrelevants relevant and thereby defeat the test. We strongly suspected that this could be done because the determinants of P300 amplitude were well-known-- low subjective probability, task relevance and/or meaningfulness-- and we knew that the methods described above for covertly increasing the meaningfulness and task relevance of irrelevant stimuli were readily implemented. This was clearly demonstrated empirically in the earlier portion of this report. In contrast, the specific *psychological/behavioral* factors of scalp distribution shapes are not well-known and have barely been studied. Even in our published demonstrations of differing profiles during deception vs truth-telling in simpler paradigms than the present one, we did not know *why* deception produced one profile, truth-telling produced another. We could but *speculate* that deceptive and truthful mind states differentially engaged neuronal systems in different ways. This is a highly abstract formulation, however, and one still has no idea how to alter a brain map of any kind in any way.

We have ourselves chosen in several previous papers and even here in the countermeasure studies to compare averaged probe P300 amplitudes to the average of responses to all irrelevants, despite the fact that this comparison inevitably entails comparing an average with relatively low numbers of single sweeps to an average based on relatively high numbers of single sweeps. These comparisons involved not profile, but simple amplitude at one site. In these studies, typically large control groups of innocent subjects or conditions has allowed us to show that there are virtually no false positives with these numerically asymmetrical comparisons in simple autobiographical oddball recognition paradigms. For the following reasons, the same arguments probably do *not* apply when the comparisons made are between *profiles*:

1) Empirically, Table 8a above showed that using comparisons of guilty/relevant items with *all* innocent (irrelevant) items afforded only 6/16 (37.5%) correct detections (from CAT) but with 2/7 (28.5%) false positives. These were based on b-p values, but if one used an either/or (bp-pp) criterion (as in Table 9) there would have been only 1 more correct detection, accompanied by 1 more false positive; the percentages just given would become 44% and 43%. Of course with the hit and false positive rates at virtually the same low level, this method is worthless. In comparison, Table 9 showed that with unconfounded comparisons within subjects, the either/or criterion yielded 73% detection against 0% false positives.

2) More theoretically, by averaging all irrelevant/innocent response profiles together, one is more likely to get a null profile to be compared with a profile for just one or two guilty (relevant) guilty act(s). It is conceivable that each stimulus type, whether guilty or innocent, whether

accused or not, could generate a profile influenced in part by the specific nature of the specific act represented. Averaging across all irrelevant acts will tend to cancel these effects in averages of many items. This would not happen with simple P300 amplitude. The P300 amplitude depends on factors quite apart from the specific meanings of evoking stimuli, namely, subjective probability and relevance. Thus, there are typically no or very small P300s to each innocent irrelevant, apart from specific meanings. Averaging them removes only noise. This means, as we saw, that the *comparison set 1* approach will work with simple amplitude at one site, but not with profile.

3) Finally, to the extent that the older simple P300 amplitude index of recognition may be used, it will definitely profit by use on a virtual site extracted with a spatial PCA.

### References

- Allen, J., Iacono, W.G. and Danielson, K.D. (1992). The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology*, 29, 504-522.
- Catell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Donchin, E., Spencer, K., & Dien (1997). The varieties of deviant experience: ERP manifestation of deviance processors in Van Boxtel, G.J.M. & Bocken, K.B.E. (Eds.), *Brain and Behavior: Past, Present, and Future*, Tilburg University Press, p. 116.
- Ellwanger, J., Rosenfeld, J.P., Sweet, J.J. & Bhatt, M. (1996). Detecting simulated amnesia for autobiographical and recently learned information using the P300 event-related potential. *International Journal of Psychophysiology*, 23, 9-23.
- Ellwanger, J., Rosenfeld, J.P., Hannkin, L.B., & Sweet, J.J. (1999). P300 as an index of recognition in a standard and difficult match-to-sample test: A model of Amnesia in normal adults. *The Clinical Neuropsychologist*, 13, 100-108.
- Fabiani, M., Gratton, G., Karis, D., & Donchin, E (1987). The definition, identification, and reliability of measurement of the P3 component of the event-related potential. In P.K. Ackles, J.R. Jennings, & M.G.H. Coles (Eds.), *Advances in psychophysiology* Vol. 2, Greenwich: JAI Press.
- Farwell, L.A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related potentials. *Psychophysiology*, 28, 531-547.
- Farwell, L.A. & Smith, S.S. (2001). Using Brain MERMER Testing to Detect Knowledge Despite Efforts to Conceal. *J. Forensic Sciences*. 46 (1), 1-9.
- Johnson, M.M., & Rosenfeld, J.P. 1992). Oddball-evoked P300-based method of deception detection in the laboratory II: Utilization of non-selective activation of relevant knowledge. *International Journal of Psychophysiology*, 12, 289-306.
- Johnson, R., Jr. (1988). The amplitude of the P300 component of the event-related potential. In P.K. Ackles, J.R. Jennings, & M.G.H. Coles (Eds.), *Advances in psychophysiology*, Vol. 2 (pp. 69-138). Greenwich, CT: JAI Press.
- Johnson, R. (1993). On the neural generators of the P300 component of the event-related potential. *Psychophysiology*, 30, 90-97.
- Lykken, D.T. (1981). *A tremor in the blood*. New York: McGraw-Hill.

- McCarthy, G. & Wood, C. (1985). Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. *Electroenceph. Clin. Neurophysiol.*, 62, 203-208.
- Miller, A.R. (1999a). P300 amplitude and topography in pseudomemory phenomena. Unpublished Doctoral Dissertation, Northwestern University, Evanston, IL, pp 11-59.
- Miller, A.R. (1999b). P300 amplitude and topography in pseudomemory phenomena. Unpublished Doctoral Dissertation, Northwestern University, Evanston, IL, pp 60-122.
- Miller, A.R., Rosenfeld, J.P., et al. (2002). P300 amplitude and topography distinguish between honest performance and feigned amnesia in an autobiographical oddball task. *J. Psychophysiology*, 16, 1-11.
- Pivik, R.T., Broughton, R., Coppola, R., Davidson, R.J., Fox, N., & Nuwer, M. (1993). Guidelines for the recording and quantitative analysis of electroencephalographic activity in research contexts. *Psychophysiology*, 30, 547-558.
- Rosenfeld, J.P., Angell, A., Johnson, M., & Qian, J. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology*, 38, 319-335.
- Rosenfeld, J.P., Cantwell, G., Nasman, V.T., Wojdacz, V., Ivanov, S., & Mazzeri, L. (1988). A modified, event-related potential-based guilty knowledge test. *International Journal of Neuroscience*, 24, 157-161.
- Rosenfeld, J.P., Ellwanger, J.W., Nolan, K., Wu, S., Bermann, & Sweet, J.J. (1999). P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. *Int. J. Psychophysiol.*, 33(1), 3-20.
- Rosenfeld, J.P., Reinhart, A.M., Bhatt, M., Ellwanger, J., Gora, K., Sekera, M., & Sweet, J. (1998). P300 Correlates of simulated amnesia on a matching-to-sample task: Topographic analyses of deception vs. truth-telling responses. *International Journal of Psychophysiology*, 28, 233-248.
- Rosenfeld, J.Peter, Rao, Archana, Soskins, M., and Miller, A.R. (2002) P300 Scalp Distribution as an Index of Deception: Control for Task Demand. *Journal of Credibility Assessment and Witness Psychology*. 3 (1) 1-22. [Internet Journal]
- Rosenfeld, J.P., Rao, A., Soskins, & Miller (submitted). Scaled P300 scalp distribution correlates of deception in an autobiographical oddball paradigm.
- Seymour, T.L., Seifert, C.M., Shafto, M.G., Mosmann, A.L., (2000). Using response time measures to assess "guilty knowledge". *Journal of Applied Psychology*, 85 (1), 30-37

- Soskins, M., Rosenfeld, J.P., & Niendam, T. (2001). The case for peak-to-peak measurement of P300 recorded at .3 hz high pass filter settings in detection of deception. *Int. J. Psychophysiology*, 40, 173-180.
- Srebro, R. (1996). A Bootstrap method to compare the shapes of two scalp fields. *Electroenceph. Clin. Neurophysiol.*, 100, 25-32.
- United States General Accounting Office (2001) *Report to Hon. Charles E. Grassley, U.S. Senate Investigative Technoques: Federal Agency Views on the Potential of "Brain Fingerprinting."* USGAO: GAO-02-22
- Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology*, 26, 208-221.

## **APPENDIX 1**

# *The Journal of Credibility Assessment and Witness Psychology*

2002, Vol. 3, No. 1, pp. 1-22

Published by the Department of Psychology of Boise State University

## **P300 Scalp Distribution as an Index of Deception: Control for Task Demand**

**J. Peter Rosenfeld, Archana Rao, Matthew Soskins,  
& Antoinette Reinhart Miller**

**Northwestern University, Department of Psychology**

Correspondence regarding this article should be addressed to Dr. J. Peter Rosenfeld, Department of Psychology and Institute for Neuroscience, Northwestern University, 2029 Sheridan Rd., Evanston, IL 60208. EMAIL: [jp-rosenfeld@northwestern.edu](mailto:jp-rosenfeld@northwestern.edu)

Copyright 2002 by the Department of Psychology of Boise State University and the Authors. Permission for non-profit electronic dissemination of this article is granted. Reproduction in hardcopy/print format for educational purposes or by non-profit organizations such as libraries and schools is permitted. For all other uses of this article, prior advance written permission is required. Send inquiries by hardcopy to: Charles R. Honts, Ph. D., Editor, *The Journal of Credibility Assessment and Witness Psychology*, Department of Psychology, Boise State University, 1910 University Drive, Boise, Idaho 83725, USA.

---

**ABSTRACT:** Participants (n=24) experienced a baseline Block 1: they saw their phone numbers presented in a series with 6 other phone numbers. They were to say "yes" to their phone numbers, "no" to others. They were asked to repeat the first 3 digits of the phone numbers aloud. In Block 2, LIE and CONTROL groups (both n=12) were formed: participants saw a series of dates (e.g., "Mar 9"), 14% of which were their birth dates. The LIE participants were asked to lie on 50% of the trials, and to repeat all stimuli aloud. The CONTROLS were to perform honestly in Block 2, and were asked to repeat all stimuli aloud, but a random half of the stimuli backwards. The aim was to equalize task demand between groups. The results were that for both scaled and unscaled P300 amplitude, there were no differences or interactions as a function of group, or block in comparisons of responses to honest, forwards-repeated stimuli ( $p > .6$ ). For pooled Block 1-Block 2 honest responses vs Block 2 dishonest responses in the LIE group, there was a main effect of response type on unscaled amplitude (lie responses < true responses,  $p < .03$ ). Conversely, there was no main effect in the CONTROL group of the forwards/backwards manipulation ( $p > .15$ ). In scaled amplitudes, there were no interactions of group or response type with site ( $p > .2$ ) in honest, forwards responses. Comparing all scaled LIE honest with dishonest responses in the LIE group yielded a significant interaction of response type x site,  $p < .02$ . Post-hoc ANOVAs, using just Cz and Pz showed a significant interaction in the LIE but not CONTROL participants. There were no P300 latency differences between groups or conditions. In an extended replication, reaction time data did not differ between LIE and CONTROL groups. The results continue to support the notion that a P300 profile, specific for deception, may be identifiable.

---

## **P300 Scalp Distribution as an Index of Deception: Control for Task Demand**

### **Introduction**

We have previously reported that in various situations, the scaled scalp distribution (profile) of P300 amplitude differs from deceptive to truth-telling conditions, (Rosenfeld, Reinhart, Bhatt, Ellwanger, Gora, Sekera, & Sweet, 1998; Rosenfeld & Ellwanger, (1999), Rosenfeld, Ellwanger, Nolan, Wu, Berman, & Sweet, 1999). Johnson (1988, 1993) has argued that when the ERP profile differs from one condition to another, this is good evidence that the two conditions involve differing neurogenerator groups.

Although one may take advantage of differential profiles for truth-tellers and liars in practical detection of deception applications, one cannot argue from such data that the liar's profile specifically represents deception. In the paradigms previously used (Rosenfeld & Ellwanger, (1999); Rosenfeld et al, (1999); Rosenfeld et al., 1998), the task demands on the liar were greater than those on the truth-teller: The latter simply had to tell the truth whereas the liar had to maintain an instructed, random-appearing, 50% (approximately) deceptive error rate, and thus also had to decide on each trial whether or not to lie. The observed differences in profile between the two groups could have represented differences in task demand as well as differences in honesty.

In the present study, we tried to construct an honest control group having task demands comparable to those of the liar group. Specifically, we used an autobiographical oddball paradigm in which participants saw a Bernoulli (randomized) series of seven, repeatedly presented dates, 14.3 percent of which were their own birth dates. In the Lie group, participants were told to respond dishonestly on a random half of the trials (of both oddball and frequent type), and to then repeat the stimuli aloud. (Only the first three letters of the month were repeated.) In the Control group, participants were told to respond aloud honestly on all trials, but to then repeat a random half of the stimuli aloud backwards, (the rest, forwards). Both groups had comparable task demands in the terms noted above, but one group responded honestly and the other dishonestly. Differing P300 profiles would not be simply attributable to differences in task demands.

In this study, there is a second set of evidence examined which bears even more directly on putative specificity of Lie profiles: On the block of trials where the Lie participants respond dishonestly on half the trials, there is the opportunity to compare the P300 profiles associated with honest and dishonest response trials. Since task demands are the same during the entire block within the Lie group, obtained profile differences would provide support for the specificity hypothesis. We looked for but failed to find such an effect earlier (Rosenfeld et al., 1999) using a different (match-to-sample) paradigm.

We note that Johnson's (1988, 1993) interpretation of the meaning of differing scalp profiles emphasizes the possibility of differing neurogenerator sets. There is another interpretation of the differing scalp profiles in two experimental conditions: It may be that the two conditions evoke different sets of components which differentially overlap the P300 which both conditions evoke in common (Donchin, Spencer, & Dien, 1997). Either interpretation implies that the brain works in a specific way during deception, and the evidence would become the first to support a specific lie response, said to be a "dream" by Lykken (1981). Such a finding would also be a step in the direction of elucidating brain systems involved in lying.

Why might one expect differing scalp distributions in Lie and Control groups if task demand is matched? We hypothesize that a participant who is lying, even though he/she was directed to do so, has some level of self-awareness on all deceptive trials; that he/she is engaging in a behavior on which society and authority figures frown. At least some participants may thus find themselves somewhat embarrassed at being observed during lies. More important, all Lie participants (and no Control participants) know they are lying as they lie, and probably engage in further lie-specific cognitions following the decision to lie as well as following the act of lying. These cognitions would pertain to knowledge of the mismatch between the true-correct answer versus the answer they produce on a lie trial. We hypothesize that the Lie condition, but not the Control condition, will generate brain activity related (at least) to

both the additional cognitions following such mismatch experiences, as well as to self awareness of deception, and that P300 profiles may reflect these differences between Lie and Control conditions.

Differences between Lie and Control groups might also be expected on the basis of the latter's additional task: backwards repetition of stimuli. A comparison restricted to profiles of Lie and Control groups during their respective specific tasks could thus be confounded by the two task effects simultaneously operating: 1) honest vs. dishonest responding and 2) backwards vs. forwards repetition. We therefore ran both groups through two blocks of trials, one (Block 1) in which all participants behaved alike in responding honestly and repeating stimuli forwards, and a second block (Block 2) in which the Lie participants lied on half the trials with forwards repetition, and the Control participants responded honestly on all trials but repeated half the stimuli in a backwards manner. Thus in each group, we could compare departures in Block 2 from the benchmark/baseline condition of Block 1.

### Method

**Participants:** The 24 participants (12 per group, 13 female, six of which were in the Lie group) were recruited from the department introductory psychology pool and were fulfilling a course requirement. All had normal or corrected vision.

**Procedure:** Following signing of consent form, instruction, and electrode attachment, participants were seated in a recliner such that a video display screen was in front of their eyes. The visual stimuli were presented on this screen every 6.0 s, a relatively long interstimulus interval required for verbal responding so as to allow the artifact associated with vocalization to dissipate prior to the subsequent trial. The trial began with the onset of pre-stimulus EEG baseline recording for 104 ms. The stimulus then appeared on the screen and endured for the remainder of the ERP recording epoch = 1944 ms (total epoch = 2048 ms). Immediately after clearance of the stimulus from the screen, the message "Please Respond" was presented and lasted 2 s. The participant was required to respond during this time.

There were two blocks of trials used in this study. In the first block (Block 1), the visual stimuli were participants' phone numbers ( $p = .14$ ) and other phone numbers ( $p = .86$ ), each repeated as many times (about 40) as the subject's phone number. Both Control and Lie participants were told to respond aloud truthfully and ordinarily in this preliminary block. The timing and parametric settings in this benchmark/baseline block were the same as in the actual test block (2) to be next detailed. In this second block (Block 2), the stimuli were the first three letters of a month, followed by a number from 1 to 31, e.g., MAR 9. Thus, birth dates could be formed. The participant then said

"yes" or "no" signifying birth date or other date, respectively, and then immediately repeated aloud the three-letter symbol of the month.

In the Control group, the participants were (in Block 2) instructed to respond honestly "yes" or "no" and to then repeat these month symbol letters aloud backwards on approximately half the trials of both types (birth date, non-birth date). They were also instructed to try giving a random, as opposed to patterned, series of forward and backward responses. We suggested to these participants that we were interested in how well people can generate random sequences of responses while doing a foreground task. We also alerted them that if the computer detected patterned responding, the experiment would be re-started.

In Block 2, the Lie group participants were instructed to simulate malingered cognitive deficit as in Rosenfeld et al. (1998), by making dishonest "errors" on both trial types about half the time in response to the "Please Respond" message. They were told to generate a random, unpatterned series of deceptive responses, since the computer controlling the experiment could discern patterns, and that they would not "beat the test" if patterned responding was discerned, and the experiment would be re-started. Immediately after their "yes" or "no" response, they were required to repeat the first three letters of the month (in the normal, forwards order). Both groups were told there would be 45 presentations of birth dates randomly interspersed among 276 presentations of other dates; i.e., six dates each repeated 45 times. This was done in order to help them score close to the 50% target rate of deceptive or backwards responses. Following the response window (2.0 s) was a second 2.0 s period of no events prior to the start of the next trial. (Verbatim instructions are available on request from the senior author.) Table 1 presents stimulus-response combinations for both groups and both blocks, with abbreviations.

**EEG recording and analysis:** EEG was recorded with Grass P511k preamplifiers with gain = 100,000, and filters set to pass signals between 0.1 and 30 Hz (3db points). Electrodes (Ag - AgCl) were attached to Fz, Cz, and Pz referenced to linked mastoids with the forehead grounded. Impedances were maintained below 5000 ohms. EOG was recorded from a bipolar pair of electrodes above and below the eye. EOG signals > 80 uV led to trial rejection and replacement. Amplified signals were led to 12-bit A/D converters (Keithley-Metrabyte) sampling at 125 Hz, and the digitized signals led to a computer for on-line sorting, averaging, and storage. The computer programs (by the senior author) also controlled stimulus presentation, and performed off-line filtering and analyses.

In the present study, P300 determination is based on a standard baseline-to-peak method: The computer searches within each participant's average ERP within stimulus, paradigm and response categories (see Table 1),

within a window which extends from 400 to 1000 ms post-stimulus for the 104 ms segment average (13 data points) which is most positive-going. From this segment average, the average of the first, pre-stimulus, 104 ms of the recording epoch is then subtracted. The difference defines unscaled P300 amplitude. The midpoint of the maximally positive segment defines P300 latency. This is a typical method of measuring P300 (Fabiani, Gratton, Karis, & Donchin, 1987).

**Table 1: Abbreviation Summary of stimulus-response combinations:**

**(a.) LIE Group**

Test Block 1 (all forward honest responses)

OD1[L]: oddball stimulus, honest response

FR1[L]: frequent stimulus, honest response

Test Block 2 (all forward responses)

OD2-TRU: oddball stimulus, honest response

OD2-LIE: oddball stimulus, dishonest response

FR2-TRU: frequent stimulus, honest response

FR2-LIE: frequent stimulus, dishonest response

**(b.) CONTROL Group**

Test Block 1 (all forward honest responses)

OD1[C] oddball stimulus, forward response

FR1[C] frequent stimulus, forward response

Test Block 2 (all honest responses)

OD2-FOW: oddball stimulus, forward response

OD2-BAC: oddball stimulus, backward response

FR2-FOW frequent stimulus, forward response

FR2-BAC frequent stimulus, backward response

The method just described is done only with Pz recordings. For the Cz and Fz sites, the temporal boundaries of the maximally positive segment at Pz are used to define the window over which P300 amplitude is calculated. This procedure is utilized to be certain that the same neural process is sampled across sites for purposes of profile construction. It is typically used by researchers who focus on scaled P300 amplitude profiles (e.g. Ruchkin, Johnson, Grafman, Canoune, & Ritter, 1992).

For group analyses, P300 latency and amplitude were based on unfiltered averages for each participant. For display, averages were digitally filtered to pass low frequencies; 3db point: 4.23 Hz. For task-by-site interactions, average P300 amplitudes within each participant were filtered and then scaled using the vector length method (McCarthy & Wood, 1985): Within each group and/or stimulus/response condition, the average Fz, Cz, and Pz values for the condition/group were squared, and the square root of the sum of the squared values was used as a denominator by which individual Fz, Cz, or Pz values within the condition/group were divided.

It is noted that analyses are performed here on both scaled and unscaled data. To look at main effects of group, stimulus type, block, response type, and scalp site on amplitude, it is appropriate to look at unscaled data (McCarthy & Wood, 1985). However, to answer questions involving interactions with site, (the major questions here) McCarthy & Wood (1985) explained the need for analysis on scaled data. What the scaling accomplishes is the removal of possible amplitude differences between conditions, which may confound amplitude distribution differences. The scaling procedure in the present study removes main effects of group, paradigm, response type, and stimulus type, and allows meaningful interpretation only of interactions involving site. Thus, as recommended by McCarthy & Wood (1985), we report analyses on both scaled and unscaled data, as appropriate. (Latency need not be scaled).

**Extended Replication:** The above procedures were repeated one year later, with one modification, on two new groups of Lie and Control subjects, (N=10, 11 respectively): Interspersed randomly among the oddball and frequent trials were 20 probe trials. On these trials, the word "Go" appeared on the computer screen and all participants were instructed to press a response button as soon as possible thereafter. This allowed us to obtain reaction time (RT) data and compare RTs between Control and Lie groups. Such information could then support our contention of equalization of task demands between groups; (RT is frequently used to assess task demand.) The probe trial stimuli appeared with the same timing as the other stimuli. Although electrodes were attached as in the original study and ERPs recorded, the ERP analysis presented is based on the original experiment. The modified replication was analyzed here only for RT data.

## Results

Note: The key quantitative results on scaled data are in sections E and F below, and in Figure 6. Other results are reported immediately below in sections A, B, C, and D.

**A. Behavioral (original study):** The mean numbers of responses in each stimulus-response category (see Table 1 for abbreviations) are shown in Table 2. There are six rows in each group and the numbers in the first row in the Lie group should correspond to those in the first row in the Control group, the second row in the Lie group with the second row in the Control group, and so on. The appropriate correspondences are close except for the fifth (second to the last) row, involving frequent stimuli (Lie = 101.58 vs. Control = 87.75). For the first four rows involving the oddball responses in both groups in both blocks, and the frequent of Block 1, there were no significant differences.

**Table 2: Average numbers ( $\pm$  SEM) of responses in each possible stimulus-response category. Table 1 and text define category abbreviations.**

Row	Category	Number
<b>Lie Group</b>		
1	OD1[L]	24.67 $\pm$ .97
2	FR1[L]	146.58 $\pm$ 5.80
3	OD2-TRU	17.25 $\pm$ .85
4	OD2-LIE	15.00 $\pm$ .90
5	FR2-LIE	101.58 $\pm$ 4.46
6	FR2-TRU	86.50 $\pm$ 4.24
<b>Control Group</b>		
1	OD1[C]	25.67 $\pm$ .99
2	FR1[C]	143.67 $\pm$ 7.16
3	OD2-FOW	15.17 $\pm$ .91
4	OD2-BAC	14.83 $\pm$ 1.28
5	FR2-BAC	87.75 $\pm$ 6.54
6	FR2-FOW	87.80 $\pm$ 5.26

There were significant effects regarding the last two rows containing frequent stimulus data, however these will not be detailed since all ERP analysis will focus only on oddball trials; P300s in many participants on frequent trials in

both groups were dubious. The present behavioral data indicate comparability between groups for oddball stimulus-response combinations; (the differences found for frequenters were small though significant).

**B. RT data (modified replication):** Average RTs to probe stimuli within each subject were averaged to yield separate group means, for each of the two blocks. For the first block in which all subjects performed in the same manner, the mean RT (+/- SD) for the Control group was 1.109S (+ .3984) and for the Lie group was 1.305S (+ .2098). On this difference,  $t(19) = 1.425$ ,  $p = .17$  (ns). In the critical second block, the differences were similar: Control = 1.02 S (+ .3925), Lie = 1.221S (+ .1927);  $t(19) = 1.47$ ,  $p = .16$  (ns). These negative data suggest that the two tasks did not impose differential demands on the two groups of subjects.

**C. ERP data: Qualitative observations in grand average ERPs:** In the first block, there should be no ERP differences between groups in response to either oddball or frequent stimuli, since both Lie and Control groups are behaving exactly alike in this block (see Table 1 and methods). Differences between groups in amplitude and latency of P300 did not, in fact, reach significance (see below).

For quality control purposes, Figure 1 shows superimposed Lie and Control grand averages for OD2-TRU and OD2-FOW trials (all honest, forwards responses in block 2). It appears that the P300 is reduced in the Lie group relative to the Control group. Figure 2 shows superimposed Lie and Control grand averages for OD2-LIE (dishonest, forwards) and OD2-BAC (honest, backwards) trials, and again, the P300s appear larger in the Control group.

Figure 3 shows superimposed OD2-TRU (honest) and OD2-LIE (dishonest) responses within the Lie group. The former set appears to have more positive P300 responses, especially at Fz and Cz. (The differences would be more obvious if we chose, in the figures, to superimpose pre-stimulus baselines, which our P300 calculation algorithm does do. We present data in figures as they really are, i.e., with random-noise related baseline shifts.) In Figure 4, comparable superimpositions are shown within the Control group: OD2-FOW (forwards) vs. OD2-BAC (backwards). In this comparison, P300 in the latter category appears slightly more positive (which, again, would be more evident with aligned baselines).

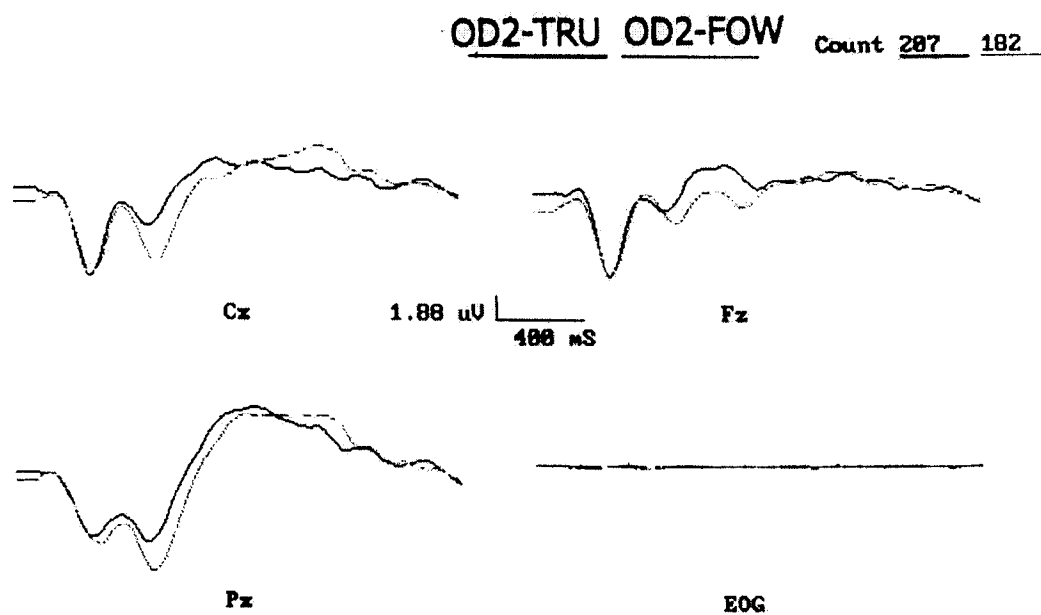


Figure 1. Superimposed grand average responses to oddballs (birthdates) in Lie (OD2-TRU) and Control (OD2-FOW) groups in the second block. Positivity is down in all ERP figures. In all ERP figures, "Count" = number of sweeps/average. Lie group (thick) and Control group (thin) are superimposed in Figs. 1-2.

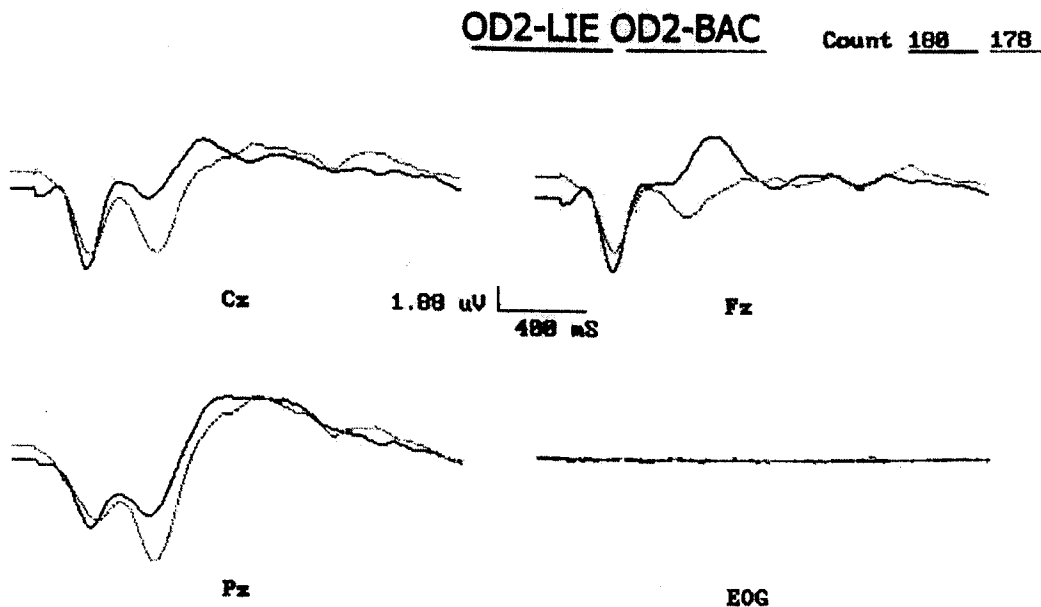


Figure 2. Superimposed oddball (birthdate) responses during lies in the Lie group (OD2-LIE) and during backwards-repetition responses in Control group (OD2-BAC), all from Block 2.

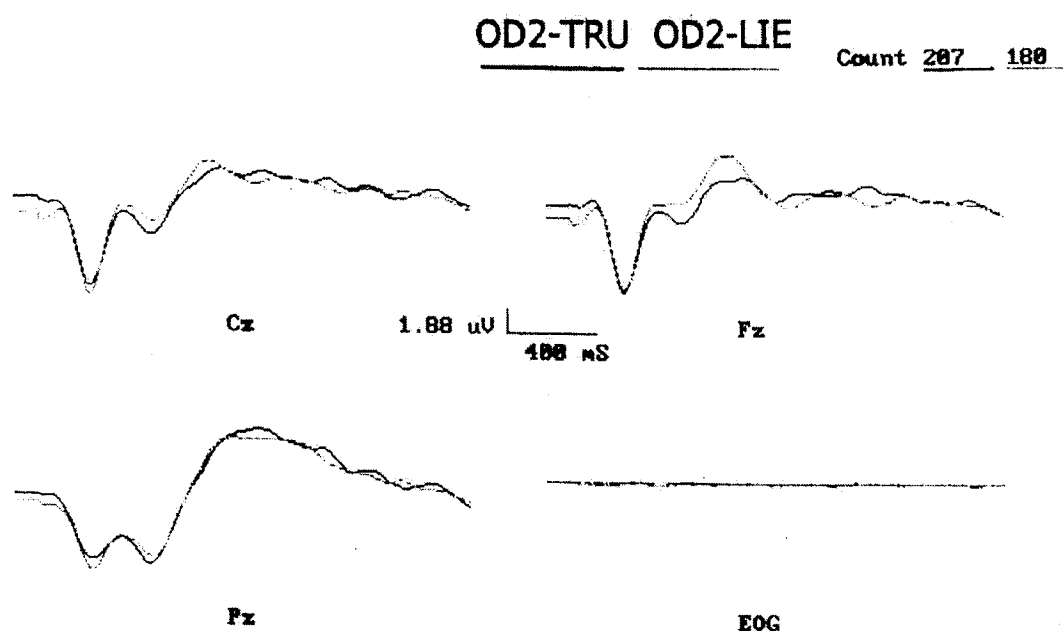


Figure 3. Superimposed honest (thick: OD2-TRU) and dishonest (thin: OD2-LIE) responses, all from LIE group in Block 2.

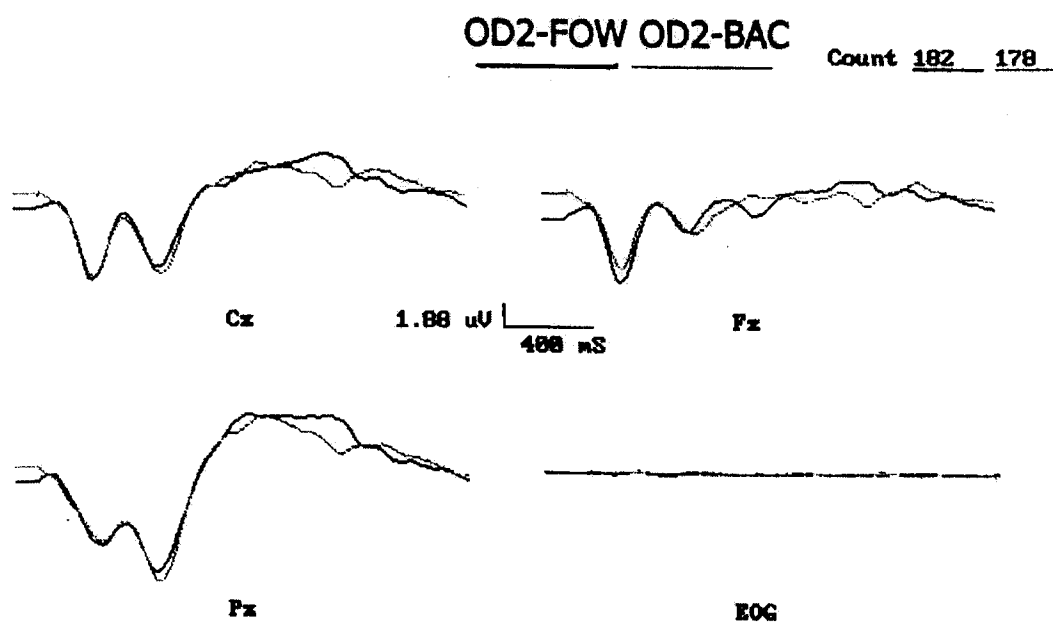


Figure 4. Superimposed forwards (thick: OD2-FOW) and backwards (thin: OD2-BAC) honest responses in Control group, Block 2.

**D. P300 amplitude data analysis: Unscaled data:** We restrict reporting of results to oddball trials, since it was frequently impossible to locate a clear P300 peak in the frequent averages within participants.

Figure 5 shows the group average, computer-determined P300 amplitude values as functions of site, group, block (1 vs 2), and stimulus-response combination. It appears that within the Lie group, there is little difference in amplitude or slope, between OD1-[L] and OD2-TRU amplitudes (both associated with honest responses), but that lying (OD2-LIE) produces a depression of amplitudes. In the Control group, the OD1[C] and OD2-FOW response curves are also aligned, and indeed do not appear to differ from comparable Lie group honest response curves just described. This is as predicted. However, in the Control group, the OD2-BAC amplitudes appear enhanced by the backward condition manipulation.

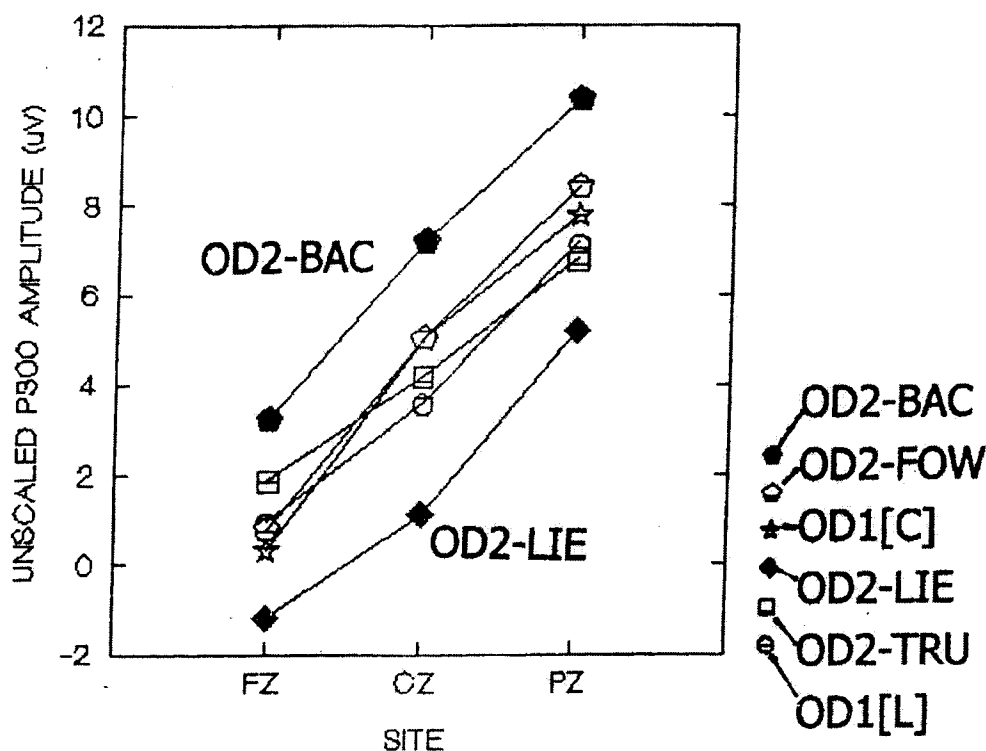


Figure 5. Averages of computer-determined, within-participant, unscaled P300 amplitudes (uV) as a function of site, paradigm, stimulus, and response type.

To obtain statistical confirmation of these effects, we first examined possible group and block differences during honest, forwards responses OD1[L], OD2-TRU, OD1[C], OD2-FOW. The sets of P300 amplitudes classified in this way were submitted to a 3-way ANOVA, with independent variables group (Lie vs. Control), site, and Block (1 vs. 2 for both groups).

The effect of group was not significant ( $p > .7$ ). Neither was the effect of Block ( $p > .6$ ). The effect of site yielded  $F(2,44) = 134.34$ ,  $pg < .001$  ( $pg$  is the Greenhouse-Geiser corrected probability in within-subject tests with  $df > 1$ . The correction is for sphericity effects. For  $df = 1$  tests, the usual  $p$ -values will be reported.) The interactions were not significant, ( $p > .2$ ), excepting the group-by-site interaction, which yielded  $F(2,44) = 4.18$ ,  $pg < .04$ , reflecting the somewhat steeper slopes for honest, forwards Control curves than for the honest, forwards Lie curves in Figure 5. (As noted in the methods, without scaling or normalization of amplitudes, all interaction effects or lack of interactions, are possibly confounded and not simply interpretable).

To get at the effects of primary interest here, we compared each of the Block 2 special response types with their respective pooled truth-telling/forwards-repeating values. (Since the 3-way ANOVA described above showed no differences between groups or block during truth-telling and forwards-repeating trials, the pooling was legitimate.) Thus we averaged OD1[L] and OD2-TRU to form OD-TRU, and we averaged the comparable Control data to form OD-FOW.

Within the Lie group, we then compared OD-TRU (honest) and OD2-LIE (dishonest) and examined site effects. The effect of site was  $F(2,22) = 89.98$ ,  $pg < .001$ . The effect of honest vs. dishonest responses was  $F(1,11) = 7.11$ ,  $p < .03$ , reflecting the lower value of averaged OD2-LIE responses in comparison with averaged OD-TRU (the pooled average of OD1[L] and OD2-TRU). The interaction of site and response type was not significant ( $p > .4$ ). In the Control group, the effect of site was  $F(2,22) = 73.36$ ,  $pg < .001$ . There was no significant effect of forwards versus backwards repetition ( $p > .2$ ), despite the appearance of such a difference in Figure 5. Neither was the interaction of response type and site ( $p > .6$ ) significant. Thus, although the dishonest response manipulation had a significant effect on unscaled P300 amplitudes in comparison with honest responses, the backwards repetition manipulation did not.

**E. P300 Amplitude analysis; scaled data: group comparisons:** In this section, we will comment only on interaction effects, since the scaling of data intentionally obviates main effects other than site effects, which are exaggerated (McCarthy & Wood, 1985). Figure 6 is the scaled equivalent of Figure 5, and shows scaled P300 amplitudes as a function of site, block, group, and response type. The figure suggests that all curves are similar except for the curve of the Lie group, during the second block, and only on dishonest response trials (OD2-LIE). We imply no interpretation of these scaled data which we simply here display (Figure 6) and describe (Ruchkin, Johnson, & Friedman, 1999).

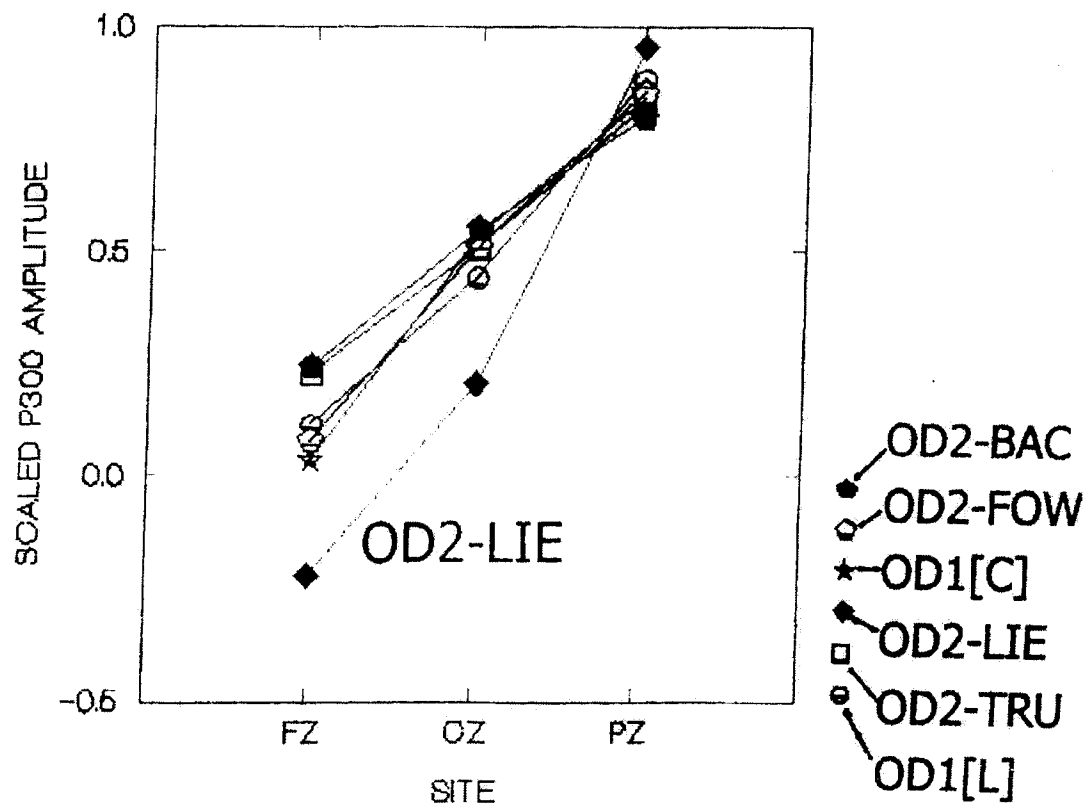


Figure 6. Averages of computer-determined, within-participant, scaled P300 amplitudes (uV) as a function of site, paradigm, stimulus, and response type.

Our statistical analysis approach with scaled data parallels the approach used with unscaled data. Thus the first analysis performed on scaled data was a 3-way ANOVA on all honest-responding, forward-repetition conditions, i.e., with independent variables: site, group, and block. The four response types separately submitted by group were OD1[L], OD2-TRU, OD1[C], and OD2-FOW. No interactions were expected, and none were found; (all  $p > .2$ ).

Next, as with unscaled data, we combined the honest, forward response trials within each group to use as a benchmark-baseline with which to compare dishonest (Lie) or backwards (Control) responses. Thus, OD-TRU is the average of OD1[L] and OD2-TRU in the Lie group; OD-FOW is the comparable average within the Control group. Within the Lie group, a 2-way ANOVA on effect of response-type (OD-TRU vs. OD2-LIE) and site yielded a significant interaction of response type-by-site;  $F(2,22) = 6.76$ ,  $pg < .02$ . Within the Control group, the comparable ANOVA on effect of OD-FOW vs. OD2-BAC with site also yielded a significant interaction;  $F(2,22) = 10.6$ ,  $pg < .001$ . This was in contrast to what is suggested in Figure 6, where the scaled curves seem all alike (especially at Cz and Pz) except for the OD2-LIE (dishonest response)

curve. It is noted (Figure 6), however, that whereas in the Lie group, the interaction shows (at Fz and Cz) a depression of OD2-LIE in comparison with OD2-TRU (honest vs. dishonest responses), in the control group, the OD2-BAC responses are (at Fz) slightly (though significantly) enhanced in comparison with the OD2-FOW curve. (These observations, again, imply no interpretation regarding relative activities or source strengths at the noted sites, but are meant simply to illustrate different kinds of interactions in Lie and Control groups; Ruchkin et al., 1999.)

We performed comparable ANOVAs, post-hoc, on data from just the Cz and Pz sites. In the Lie group, OD-TRU vs. OD2-LIE interacted with site,  $F(1,11) = 24.32$ ,  $p < .001$ . However, in the Control group, OD-FOW vs. OD2-BAC did not interact with site ( $p > .15$ ); neither did OD1-FOW vs. OD2-BAC ( $p > .1$ ).

**F. P300 Scaled Amplitude Analysis: Within Lie Group:** The major comparison in this study is of the honest and dishonest response trials in Block 2 within the Lie group (OD2-TRU vs. OD2-LIE). This is because the task demands in the Lie group should be constant over trials within the block. A 2-way ANOVA on response type (honest/dishonest) and site did yield an interaction:  $F(2,22) = 7.2$ ,  $pg < .02$ , as is evident also in Figure 6.

**G. Latency Effects:** Table 3 shows the Pz latencies of P300 for oddball responses in the two groups, segregated by response type. The Control group latencies are slightly greater than those of the Lie group (although the largest difference in row 1 of the table occurs prior to the group-generating manipulation). For both groups responding honestly and with forwards repetition in both paradigms, a 2-way ANOVA was performed on oddball latencies, with independent variables group and response type. There were no significant effects for group ( $p > .2$ ), response type ( $p > .5$ ) or interaction ( $p > .4$ ).

Another 2-way ANOVA was performed on Pz latencies involving group and honest, dishonest, forwards, and backwards response types. Again there were no significant effects of group ( $p > .4$ ), response type ( $p > .6$ ), or interaction ( $p > .6$ ). The present manipulations had no effects on P300 latencies, suggesting that stimulus processing task demands for the two groups did not differ, inasmuch as P300 latency has been associated with stimulus evaluation time (Fabiani et al., 1987; Johnson, 1988).

Table 3: P300 Pz Latencies  $\pm$  SD

Lie Group		Control Group	
Response Type	Latency (ms)	Response Type	Latency (ms)
OD1[L]	516 + 34.9	OD1[C]	550 +/- 53.9
OD2-TRU	518 + 80.2	OD2-FOW	528 +/- 47.3
OD2-LIE	518 + 49.6	OD2-BAC	539 +/- 44.8

### Discussion

We have shown previously (Rosenfeld et al, 1998; Rosenfeld et al., 1999) that the scaled scalp distributions (profiles) of P300 amplitude in deception conditions differ from those seen in simple truth-telling conditions. Since the scaled scalp amplitude distribution is independent of amplitude itself (McCarthy & Wood, 1985; Johnson, 1988, 1993), it may well be the case that profile can become another brain-wave-based channel (dependent measure) which could be used in practical detection of deception situations. There have now been several demonstrations that P300 amplitude, itself, can be so utilized; (e.g., Rosenfeld, Cantwell, Nasman, Wojdac, Ivanov, & Mazzeri, 1988, Rosenfeld, Angell, Johnson, & Qian, 1991, Ellwanger, Rosenfeld, Sweet, & Bhat, 1996, Farwell & Donchin, 1991; Allen & Iacono, 1992.)

One could not say, however, on the basis of previous studies, that the profile seen in deceptive conditions represented neural activity specific to deception, itself, since, as reviewed in the introduction, deceptive and truth-telling conditions previously utilized also differed in task demand: the truth-teller had only to do his/her best on a simple task whereas the deceiver had to (additionally) keep track of his/her deception rate, and decide on each trial whether or not to lie.

The present study was designed to address these considerations in 2 ways: (1) allowing comparison of profiles between two groups (Lie and Control) in which we attempted to equalize task demand to the maximum possible extent, and (2) allowing comparison within the Lie group of profiles associated with honest versus dishonest response trials. Differing profiles in dishonest versus honest conditions would suggest different neurogenerator sets associated with each condition (Johnson, 1993; McCarthy & Wood, 1985). It may also be that the two conditions evoke different sets of components which

differentially overlap the P300 which both conditions evoke in common (Donchin et al, 1997). In either case, however, the differing profiles indicate differing modes of brain function in each condition.

In fact, we found (Results, section F.) that scaled profiles differed in Lie group members during honest versus dishonest response trials. Since the task demand on the Lie group members was the same throughout the second paradigm task (i.e. during honest and dishonest trials), it is suggested that the significant interaction of response type (honest vs dishonest) by site provides evidence of differential modes of brain operation during the two kinds of trials, and that this effect is not confounded by task demand differences.

The Control group, like the Lie group also had to make a decision on each trial (whether or not to repeat a stimulus backwards), and had to track the same ratio of the two kinds of available responses (50-50). When scaled amplitude data from all three sites (Fz, Cz, Pz) were analyzed, this group also showed an interaction of site and response type (honest forwards repetition vs honest backwards repetition). However, the nature of the change from the forward repetition condition in the Control group was different than that seen in the Lie group. Indeed, if one considered only the Cz and Pz sites, then only the Lie group showed an interaction effect in the response type manipulation (response type x site) whereas the Control group showed no (response type x site) significant interaction. Similarly, in unscaled data from all three sites, significant main effects on amplitude were seen only in response to the honesty manipulation and not in response to the forwards vs backwards repetition manipulation. Thus the honesty-dishonesty manipulation had greater effects than the forwards-backwards manipulation (on unscaled Fz, Cz, Pz amplitudes, and on scaled profiles at Cz and Pz) in this study.

Further evidence that group differences are not attributable to stimulus complexity aspects of task demand differences comes from the latency data: The P300 latencies did not differ between Lie and Control groups. Increases in task complexity involving greater stimulus processing demand from one condition to another are usually reported to increase P300 latency (and to decrease amplitude; Johnson, 1988).

It is also the case that in a modified replication of the present experiment in which probe stimuli were randomly inserted in place of date and number stimuli, there were no differences in reaction time to these probe stimuli between Lie and Control groups. This was further evidence of the comparability of task demand in these groups. We could not look at RTs to the other stimuli (as is often customary) because of the delayed response requirement necessitated by the need to avoid vocalization artifact. The probe stimuli, however, appeared in exactly the same time slots as did the other stimuli. They were more rare and when presented, were probably unexpected, as subjects most likely anticipated presentation of dates.

It is not surprising that in scaled profile data, the Control and Lie groups had differing profiles in Block 2 in comparison with their respective benchmarks. The two tasks are quite different in two ways, involving 1) honest (Control) versus dishonest (Lie) responses, and 2) trials with forward (Lie) versus backward repetition (Control). One could not say with certainty that by themselves, these differing profiles are due to honesty differences, repetition direction differences, or both. This is why we also used a first block with all participants responding honestly with forward repetition of stimuli. Since these profile data did not differ from the honest/forward repetition data in the second block, we pooled, within each group, the honest/forward response data from both blocks and used them as baseline/benchmarks with which to compare dishonest response profiles in the Lie group and backwards response profiles in the Control group. The manipulations within each group produced different scaled profile effects, in terms of shifts from the benchmarks as noted above, and we would attribute the effect in the Lie group to effects of deception.

This is consistent with the finding of different profiles for honest and dishonest responses within the second block of the Lie group, where within one block, different profiles were obtained. These effects might be attributable to deception specifically, since, as noted above, these Lie participants were all treated alike and the only difference between the cognitive states of Lie participants on trials involving honest vs. deceptive responses is this difference in response selection.

It is noted (Figure 6) that in the Lie group, the scaled OD2-LIE (dishonest response) curve is downshifted at Cz and Fz and upshifted at Pz relative to both the honest condition of Block 1 (OD1[L]) as well as to the honest response trials of Block 2 (OD2-TRU). It is also downshifted in comparison with all Control group curves at Fz and Cz, and upshifted at Pz. (We do not here intend to interpret the interactions on scaled data in terms of loci of cortical activity responsible for the interactions, as noted in the Results section, but only mean to describe unique features of the interaction in the Lie group.) These interactions strongly suggest that the lie response has a unique effect on brain operation. The fact that unscaled amplitudes are uniquely reduced in the Lie group during dishonest responses also supports a unique attribute related specifically with dishonest responses.

A question may be raised here regarding ecological validity. Our Lie subjects were not, in fact, lying in the way people do in the field. In our instructions to them, however, we repeatedly reminded them that when they would respond as if they were making errors, that in fact, they would know very well that these were not genuine errors, but lies. (One subject actually refused to complete the study at this point and was released.) Nevertheless, it remains a limitation here that the subjects were executing directed rather than voluntary lies.

It was essential, in the design of this study, that there be no differences among the P300s associated with both blocks and groups during the honest responses. This requirement was mandated by our plan to pool honest, forwards responses so as to generate benchmark/baselines as described above. However, we also had application issues in mind: In any anticipated uses of these methods with real suspects in the field, it may be essential to have data from a control/baseline session, in which the suspect is known to be responding truthfully, with which to compare, in the same subject, data obtained during a test session in which the subject's (dis)honesty is to be ascertained. The present results in the Lie group which showed no differences between P300 distributions associated with truthful responses from both the first and second blocks, but differences between pooled truthful responses and dishonest responses, suggest that it should be possible to develop procedures, based on current group results, for future intraindividual diagnosis.

There is another implication regarding the data obtained from both groups during honest, forwards responding: One might have predicted differences between data sets obtained from the two blocks during honest, forwards responding on the basis of the fact that the first block utilized phone numbers as stimuli, whereas the second block utilized (birth) dates. A participant might have been expected to show different scaled amplitude profiles to these two kinds of stimuli on the basis of different cognitive processing of the two classes. Such differences were not observed. (Of course, such differences might be seen in data from other scalp sites.) This negative outcome suggests that the specific nature of the stimulus does not play a significant role in determination of profile shape in the present context: Rather, an autobiographical oddball stimulus yields a typical  $P_z > C_z > F_z$  profile which does not differ as a function of the specific nature of the stimulus, so long as an honest response occurs to the stimulus. Dishonest responses, however, affect the profile. We could have counterbalanced across participants the order of stimulus class used in the present study in order to control (unobtained) effects of differing stimulus classes. We chose not to counterbalance because while this counterbalanced design would have been easily implemented in the present laboratory analog, it would appear to present major problems in intraindividual field tests.

## References

- Allen, J., Iacono, W.G. and Danielson, K.D. (1992). The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology*, 29, 504-522.
- Donchin, E., Kramer, A., & Wickens, C. 1986). Applications of brain event-related potentials to problems in engineering psychology. In M. Coles, S. Porges and E. Donchin (Eds.), *Psychophysiology: systems, processes and applications*. New York: Guilford.
- Donchin, E., Spencer, K., & Dien (1997). The varieties of deviant experience: ERP manifestation of deviance processors in Van Boxtel, G.J.M. & Bocken, K.B.E. (Eds.), *Brain and Behavior: Past, Present, and Future*, Tilburg University Press, p. 116.
- Fabiani, M., Gratton, G., Karis, D., & Donchin, E (1987). The definition, identification, and reliability of measurement of the P3 component of the event-related potential. In P.K. Ackles, J.R. Jennings, & M.G.H. Coles (Eds.), *Advances in psychophysiology* Vol. 2, Greenwich: JAI Press.
- Ellwanger, J., Rosenfeld, J.P., Sweet, J.J. & Bhatt, M. (1996). Detecting simulated amnesia for autobiographical and recently learned information using the P300 event-related potential. *International Journal of Psychophysiology*, 23, 9-23.
- Farwell, L.A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("lie detection") with event-related potentials. *Psychophysiology*, 28, 531-547.
- Johnson, R., Jr. (1988). The amplitude of the P300 component of the event-related potential. in P.K. Ackles, J.R. Jennings, & M.G.H. Coles (Eds.), *Advances in psychophysiology* Vol. 2 (pp. 69-138). Greenwich, Ct: JAI Press.

- Johnson, R. (1993). On the neural generators of the P300 component of the event-related potential. *Psychophysiology*, 30, 90-97.
- Kramer, A.F., Sirevaag, E.J., & Braune, R. (1987). A psychological assessment of operator workload during simulated flight missions. *Human Factors*, 29(2), 145-160.
- Lykken, D.T. (1981). *A tremor in the blood*. New York: McGraw-Hill.
- McCarthy, G. & Wood, C. (1985). Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. *Electroenceph. Clin. Neurophysiol.*, 62, 203-208.
- Rosenfeld, J.P., Angell, A., Johnson, M., & Qian, J. (1991). An ERP-based, control-question lie detector analog: Algorithms for discriminating effects within individuals' average waveforms. *Psychophysiology*, 38, 319-335.
- Rosenfeld, J.P., Cantwell, G., Nasman, V.T., Wojdacz, V., Ivanov, S., & Mazzeri, L. (1988). A modified, event-related potential-based guilty knowledge test. *International Journal of Neuroscience*, 24, 157-161.
- Rosenfeld, J.P., Reinhart, A.M., Bhatt, M., Ellwanger, J., Gora, K., Sekera, M., & Sweet, J. (1998). P300 Correlates of simulated amnesia on a matching-to-sample task: Topographic analyses of deception vs. truth-telling responses. *International Journal of Psychophysiology*, 28, 233-248.
- Rosenfeld, J.P. & Ellwanger, J.W. (1999). Cognitive Psychophysiology in Detection of Malingered cognitive deficit. *Forensic Neuropsychology: Fundamentals and Practice*, J.J. Sweet (Ed.), Lisse, Netherlands: Swets & Zeitlinger.
- Rosenfeld, J.P., Ellwanger, J.W., Nolan, K., Wu, S., Bermann, & Sweet, J.J. (1999). P300 scalp amplitude distribution as an index of deception in a simulated cognitive deficit model. *Int. J. Psychophysiol.*, 33(1), 3-20.

Ruchkin, D.S., Johnson, R., Grafman, J., Canoune, H., & Ritter, W. (1992). Distinctions and similarities among working memory processes: an event-related potential study. *Cognitive Brain Research*, 1, 53-66.

Ruchkin, D.S., Johnson, R., & Friedman, D. (1999). Scaling is necessary when making comparisons between shapes of event-related potential topography. *Psychophysiology*, 36, 832-834.

**Article submitted for publication: 14 January 2002**

**Revision submitted: 22 May 2002**

**Article accepted for publication: 8 June 2002**

## APPENDIX 2

**scenario: 1 steal ring**

probe	target	irrelevants
1 ring	watch	bracelet, chain, necklace, broach
2 drawer	desk	cabinet, cupboard, box safe
3 red	gold	green, white, purple, yellow
4 Jim	Simon	Charles, Eric, Stuart, Nick
5 pocket	pants	shoe, wallet, sock glove
6 cow	goat	pig, chicken, horse, rhino

**scenario 2: steal grade sheet**

probe	target	irrelevants
1 gradesheet	roster	transcript, diploma, book file
2 wall	floor	table, chair, shelf, computer
3 blue	brown	pink, orange, black, gray
4 Rosenfled	Peters	Uttal, Paller, Revelle, Bailey
5 folder	backpack	briefcase, suitcase, binder, envelope
6 cat	dog	mouse, rat, money, goose

For both scenarios:

1 is item stolen, 2 is where is taken from, 3 is background color, 4 is whose it is, 5 is where it is put, 6 is operation name.

APPENDIX 2

**In the last five years, I have smoked marijuana week' .**

**In the last five years, I have not smoked marijuana weekly.**

**In the last five years, I have stolen school records.**

**In the last five years, I have not stolen school records.**

**In the last five years, I have used a fake ID.**

**In the last five years, I have not used a fake ID.**

**In the last five years, I have stolen a friend's money.**

**In the last five years, I have not stolen a friend's money.**

**In the last five years, I have plagiarized a paper.**

**In the last five years, I have not plagiarized a paper.**

**In the last five years, I have stolen a bicycle.**

**In the last five years. I have not stolen a bicycle.**

**In the last five years, I have robbed a bank.**

**In the last five years, I have not robbed a bank.**

**In the last five years, I have broken a store window.**

**In the last five vears, I have not broken a store window.**

## APPENDIX 2

For this sub-test you are going to see a series two-word phrases of anti-social acts on the screen in front of you. In addition, there will also be a control item “lie test” which is a shortened form of “taken lie test”. Each phrase will be followed by a message that reads, “respond”. When you see that message you will press one of the two buttons in front of you.

In this test it is your objective to appear innocent of all of the anti-social items, so you will press the left button marked with an “N” in response to all of them when the “respond” message appears. In response to the control question “lie test” (the only item we know you have committed you are doing it right now), you are to press the right button marked with a “Y” when the respond message appears.

This test will take approximately twenty minutes. It is important that you remain as motionless and relaxed as comfortably possible, press the “Y” button only for “lie test”, press the “N” button for all of the other acts, and respond when the “respond” message appears but not before or after.

If you have any questions, please ask the experimenter, however it is important that you do not inform the experimenter of your guilt or innocence of any of the anti-social acts.

## APPENDIX 2

In the past five years, have you:

**Smoked marijuana weekly?**      **Yes**\_\_\_\_      **No**\_\_\_\_

**Stolen school records?**      **Yes**\_\_\_\_      **No**\_\_\_\_

**Used a fake ID?**      **Yes**\_\_\_\_      **No**\_\_\_\_

**Stolen a friend's money?**      **Yes**\_\_\_\_      **No**\_\_\_\_

**Plagiarized a paper?**      **Yes**\_\_\_\_      **No**\_\_\_\_

**Stolen a bicycle?**      **Yes**\_\_\_\_      **No**\_\_\_\_

**Robbed a bank?**      **Yes**\_\_\_\_      **No**\_\_\_\_

**Broken a store window?**      **Yes**\_\_\_\_      **No**\_\_\_\_

## APPENDIX 2

**Northwestern University : Consent Form****Principal Investigator : J. P. Rosenfeld, Department of Psychology****EVOKED BRAIN POTENTIALS AND RECALL CONSENT FORM : B**

**INTRODUCTION:** You are being asked to participate in a research study designed to increase understanding of how the brain represents attention and memory.

**PROCEDURE:** Thirty-two electrodes will be attached with this "shower like" cap and paste to my scalp and face for the purpose of brain waves recording. The paste will be washed out at the conclusion of the experiment. No electric current is passed into the scalp; on the contrary my brain's electricity is recorded.

You will be asked to respond on a list to personal questions about behavior and integrity, although no record of my answer is kept and the responses will remain anonymous. You may withdraw from the study at any time without prejudice or penalty. One of the experimenters will observe you throughout the study, and that you are free to inquire at any time about any aspect of the study.

**RISKS:** Ordinarily, the only risks involved in participation is possible irritation from the paste used to attach the electrodes to the head. However, if you have used Retin-A on the face in the past six months. You should inform the experimenter, and you must decline to have electrodes placed on or near such areas. If there is doubt, you must decline to serve in the experiment. There is a definite risk of skin damage if electrodes are placed over skin areas which have recently been treated with Retin-A, or related products.

**BENEFITS:** The study may increase out understanding of how the brain represents memory and mechanisms of attention and memory disorders. You will also receive credit toward your 110 requirement. You will learn about Psychology lab research.

**SUBJECTS' RIGHTS:** Participation in this study is voluntary. You may withdraw at any time and still receive appropriate credit. Refusal to participate or withdraw will not affect school standing. Any questions about the study will be answered by Dr. J. P. Rosenfeld, (847) 491-3629. Questions about research subjects' rights may be directed to the Office for the Protection of Research Subject, (312) 503-9338.

**CONSENT:** I agree to participate in the research study outlined above.

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

\_\_\_\_\_  
Print Name

Northwestern University  
Institutional Review Board  
Approval Date 11-17-2000  
Approval Expires 11-17-2001